(2)

ON THE CONTROL OF AUTOMATIC PROCESSES: A PARALLED DISTRIBUTED PROCESSING ACCOUNT OF THE STROOP EFFECT

Technical Report AIP - 132

Jonathan D. Cohen
Kevin Dunbar &
James L. McClelland
Department of Psychology

The Artificial Intelligence and Psychology Project

Departments of
Computer Science and Psychology
Carnegie Mellon University

Learning Research and Development Center
University of Pittsburgh

(2)

ON THE CONTROL OF AUTOMATIC PROCESSES: A PARALLED DISTRIBUTED PROCESSING ACCOUNT OF THE STROOP EFFECT

Technical Report AIP - 132

Jonathan D. Cohen
Kevin Dunbar &
James L. McClelland
Department of Psychology

Carnegie Mellon University
Pittsburgh, PA 15213 U.S.A.

November 22, 1989

This research was partially supported by the Computer Science Division, Office of Naval Research, under contract number N00014-86-K-0678. Reproduction in whole or in part is permitted for any purpose of the United States Government. Approved for public release; distribution unlimited.

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release; Distribution unlimited
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) AIP - 132	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION Carnegie Mellon University 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION Computer Sciences Division Office of Naval Research (Code 1133)
6c. ADDRESS (City, State, and ZIP Code) Department of Psychology Pittsburgh, PA 15213	7b. ADDRESS (City, State, and ZIP Code) 800 N. Quincy Street Arlington, VA 22217-5000
8a. NAME OF FUNDING / SPONSORING ORGANIZATION Same as Monitoring Organization 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER N00014-86-K-0678
8c. ADDRESS (City, State, and ZIP Code)	10. SOURCE OF FUNDING NUMBERS p40005ub201/7-4-86 PROGRAM ELEMENT NO: N/A; PROJECT NO.: N/A; TASK NO: N/A; WORK UNIT ACCESSION NO: N/A
11. TITLE (Include Security Classification) On the control of automatic processes: A paralled distributed processing account of the Stroop effect	
12. PERSONAL AUTHOR(S) Jonathan D. Cohen, Kevin Dunbar, & James L. McClelland	
13a. TYPE OF REPORT Technical 13b. TIME COVERED FROM 86Sept15 TO 91Sept14	14. DATE OF REPORT (Year, Month, Day) 1989NOV22 15. PAGE COUNT 75
16. SUPPLEMENTARY NOTATION In Psychological Review, July 1990.	
17. COSATI CODES FIELD / GROUP / SUB-GROUP	18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number) automaticity cognitive psychology modelling Stroop task
19. ABSTRACT (Continue on reverse if necessary and identify by block number) SEE REVERSE SIDE	
20. DISTRIBUTION / AVAILABILITY OF ABSTRACT ☐ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION
22a. NAME OF RESPONSIBLE INDIVIDUAL Dr. Alan L. Meyrowitz	22b. TELEPHONE (Include Area Code) (202) 696-4302 22c. OFFICE SYMBOL N00014

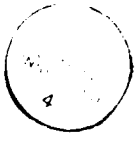
DD FORM 1473, 84 MAR 83 APR edition may be used until exhausted. All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE

Unclassified

Abstract

A growing body of evidence suggests that traditional views of automaticity are in need of revision. For example, automaticity has often been treated as an all-or-none phenomenon, and traditional theories have held that automatic processes are independent of attention. Yet recent empirical data suggest that automatic processes are continuous, and furthermore are subject to attentional control. In this paper we present a model of attention which addresses these issues. Using a parallel distributed processing framework we propose that the attributes of automaticity depend upon the strength of a processing pathway and that strength increases with training. Using the Stroop effect as an example, we show how automatic processes are continuous and emerge gradually with practice. Specifically, we present a computational model of the Stroop task which simulates the time course of processing as well as the effects of learning. This was accomplished by combining the cascade mechanism described by McClelland (1979) with the back propagation learning algorithm (Rumelhart, Hinton, & Williams, 1986). The model is able to simulate performance in the standard Stroop task, as well as aspects of performance in variants of this task which manipulate SOA, response set, and degree of practice. In the discussion we contrast our model with other models, and indicate how it relates to many of the central issues in the literature on attention, automaticity, and interference.



Psychological Review, in press 11/22/89

On the Control of Automatic Processes: A Parallel Distributed Processing Account of the Stroop Effect

Jonathan D. Cohen
Carnegie Mellon University,
University of Pittsburgh
and
Stanford University

Kevin Dunbar
McGill University

and

James L. McClelland
Carnegie Mellon University

Correspondence concerning this paper should be addressed to:

Jonathan D. Cohen
Department of Psychology
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

A growing body of evidence suggests that traditional views of automaticity are in need of revision. For example, automaticity has often been treated as an all-or-none phenomenon, and traditional theories have held that automatic processes are independent of attention. Yet recent empirical data suggest that automatic processes are continuous, and furthermore are subject to attentional control. In this paper we present a model of attention which addresses these issues. Using a parallel distributed processing framework we propose that the attributes of automaticity depend upon the strength of a processing pathway and that strength increases with training. Using the Stroop effect as an example, we show how automatic processes are continuous and emerge gradually with practice. Specifically, we present a computational model of the Stroop task which simulates the time course of processing as well as the effects of learning. This was accomplished by combining the cascade mechanism described by McClelland (1979) with the back propagation learning algorithm (Rumelhart, Hinton, & Williams, 1986). The model is able to simulate performance in the standard Stroop task, as well as aspects of performance in variants of this task which manipulate SOA, response set, and degree of practice. In the discussion we contrast our model with other models, and indicate how it relates to many of the central issues in the literature on attention, automaticity, and interference.

Introduction

The nature of attention has been one of the central concerns of experimental psychology since its inception (e.g., Cattell, 1886; Pillsbury 1908). James (1890) emphasized the selective aspects of attention and regarded attention as a process of "taking possession by the mind, in clear and vivid form, of one out of what seems several simultaneously possible objects or trains of thought" (p.403). Others, such as Moray (1969) and Posner (1975), have noted that attention is also a heightened state of arousal, and that there appears to be a limited pool of attention available for cognitive processes. Posner and Snyder (1975) and Shiffrin and Schneider (1977) have provided accounts of attention that integrate these aspects of attention and emphasize that attention is intimately tied to learning. These accounts focus on two types of cognitive processes — controlled and automatic. Controlled processes are voluntary, require attention, and are relatively slow, whereas automatic processes are fast and do not require attention for their execution. Performance of novel tasks is typically considered to rely on controlled processing; however, with extensive practice performance of some tasks can become automatic[1] (e.g., LaBerge & Samuels, 1974; Logan, 1979; Posner & Snyder, 1975; Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977).

Many tasks have been used to examine the nature of attention and automaticity. Perhaps the most extensively studied tasks have been the search tasks of Shiffrin and Schneider (1977; Schneider & Shiffrin, 1977), priming tasks (e.g., Neely, 1977), and the Stroop task (Stroop,1935). The interpretation of such studies has often relied on the assumption that automaticity is an all or none phenomenon. However, recent research has begun to question this assumption (e.g., Kahneman & Henik, 1981; MacLeod & Dunbar, 1988).

---

[1] Some authors have argued that certain automatic processes are innate. For example, Hasher and Zacks (1979) argue that the encoding of event frequency is an automatic process and that it is innate. In this paper, however, our focus is on processes that become automatic after extensive practice at a task.

An alternative conception is that automaticity is a matter of degree. For example, Kahneman and Treisman (1984) have suggested that processes may differ in the extent to which they rely on attention, and MacLeod and Dunbar (1988) have presented data which indicate that the attributes of automaticity develop gradually with practice. As yet, however, there is no explicit account of the mechanisms underlying automaticity that can explain both its gradual development with practice and its relation to selective attention. The purpose of this paper is to provide such an account.

We will begin by illustrating the relationship between attention and automaticity — as it is commonly construed — in the context of the Stroop interference task. We will show how previous attempts to explain the Stroop effect point to significant gaps in our understanding of this basic phenomenon. We will then describe a theoretical framework in which automaticity can be viewed as a continuous phenomenon that varies with practice, and which specifies the relationship between automaticity and attentional control in terms of specific information processing mechanisms. The main body of the paper will describe a simulation model that applies this theoretical framework to performance in the Stroop task.

The Stroop Task

The effects observed in the Stroop task provide a clear illustration of our capacity for selective attention, and the ability of some stimuli to escape attentional control. In this task, subjects are asked to respond to stimuli which vary in two dimensions, one of which they must ignore. In the classic version of the task, subjects are shown words written in different colored inks. When the task is to read the word, subjects are effective in ignoring the color of the ink, as evidenced by the fact that ink color has no influence on word reading time. However, when the task is to name the *ink color*, they are unable to suppress the effects of word form. If the word conflicts with the ink color (e.g., GREEN in red ink[1]) they are consistently slower to respond (i.e., say "red") than for control stimuli (e.g.,

---

[1] Throughout this paper, references to word stimuli will appear in upper case (e.g., RED), references to color stimuli will appear in lower case (red), and references to potential responses will appear in quotation marks ("red").

a row of XXXX's printed in red ink); and they are faster if the word agrees with the ink color (e.g., RED in red ink). Subjects are also slower overall at color naming than at word reading, suggesting that color naming is a less practiced task. These effects are highly robust, and similar findings have been observed in a diversity of paradigms using a wide variety of different stimuli (for reviews, see Dyer, 1973 and MacLeod, 1989). The Stroop effect illustrates a fundamental aspect of attention: we are able to ignore some features of the environment but not others.

The simplest explanation for the Stroop effect is that the relevant difference between color naming and word reading is their speed of processing. Indeed, subjects are consistently faster at reading words than naming colors. Because of this, it is often assumed that the word arrives at the response stage of processing before color information. If the word concurs with the color, this will lead to facilitation of the color naming response; if the word conflicts, its influence must be "overcome" in order to generate the correct response, leading to a longer response time for (i.e., interference with) the color naming process. Since color information arrives at the response stage after the word information, it has no effect on the word reading process.

However, if speed of processing is the only relevant variable, then it should be possible to make color information conflict with word reading, by presenting color information long enough before the onset of the word. In fact, this does not work. Glaser and Glaser (1982) varied the stimulus onset asynchrony (SOA) of a color patch and a color word,[1] and found no interference of the color patch on word reading even when the color preceded the word by as much as 400 msec. This result indicates that the relative finishing time of the two processes is not the sole determinant of interference effects.

A more general approach to explaining Stroop-like effects has been to consider the role of attention in processing. This approach draws on the distinction between automatic and controlled processes (Cattell, 1886; Posner & Snyder, 1975; Shiffrin & Schneider, 1977). Automatic processes are fast, do not require attention for their execution, and therefore can

---

[1] As we will discuss below, the Stroop effect can still be observed even when the two stimulus dimensions are physically disjoint.

occur involuntarily. In contrast, controlled processes are relatively slow, require attention, and therefore are under voluntary control. From this point of view, the results of an automatic process are more likely to escape our attempts at selective attention than are those of a controlled process.

Posner and Snyder applied the distinction between controlled and automatic processes directly to the Stroop task, by making the following three assumptions: 1) word reading is automatic; 2) color naming is controlled; and 3) if the outputs of any two processes conflict, one of the two processes will be slowed down. In this view, the finding that word reading is faster than color naming follows from the relatively greater speed of automatic processes. The finding that ink color has no effect on word processing follows from the assumption that color naming is controlled and therefore voluntary; so, it will not occur when the task is to ignore the color and read the word. The finding that a conflicting word interferes with color naming follows from the automaticity (i.e., involuntary nature) of word reading, and the assumption that conflicting outputs slow responding.

This interpretation of the Stroop task exemplifies a general method that has been used for assessing the automaticity of two arbitrary processes A and C, based on their speed of processing and the pattern of interference effects they exhibit. If A is faster than C, and if A interferes with C but C does not interfere with A, then A is automatic and C is controlled. Of course, this reasoning requires that processes A and C are in some sense comparable in intrinsic difficulty and number of processing stages.

This method for identifying processes as automatic or controlled has gained wide acceptance. However, evidence from a recent series of experiments conducted by MacLeod and Dunbar (1988) suggests that this may not be an adequate characterization of the processes involved in the Stroop task. They taught subjects to use color words as names for arbitrary shapes that actually appeared in a neutral color. After 288 trials (72 trials/stimulus), subjects could perform this "shape naming" task without difficulty. At this point, the effect that ink color had on shape naming was tested, by presenting subjects with conflict and congruent stimuli (i.e., shapes colored to conflict or agree with their assigned names). Ink color produced large interference and facilitation effects. However, when the task was reversed, and subjects were asked to state the color of the ink in which the shapes appeared (the color naming task), congruity of the shape name had no effect. They also noted that reaction times for the shape naming task (control condition) were slower than were those for the standard color naming task (control condition).

MacLeod and Dunbar's results are incompatible with the explanation of the Stroop task in terms of controlled versus automatic processing. That is, according to standard reasoning since a) color naming is slower than word reading, b) color naming is influenced by word information, while c) ink color does not influence word reading, it is assumed color naming must be *controlled*. Yet, in MacLeod and Dunbar's experiment color naming reversed roles. That is, a) color naming was faster than shape naming, b) color naming was not affected by shape names, yet c) ink color did interfere with (and facilitate) shape naming. If we treat automaticity as dichotomous, we must conclude from these findings that color naming is *automatic*.

One way of accounting for these data — rather than trying to dichotomize processes as controlled or automatic — is to suppose that tasks such as word reading, color naming and shape naming lie along a continuum. This is suggested by their relative speeds of performance and by the pattern of interference effects that exist among these tasks. Thus, word reading is faster than and is able to interfere with color naming, while color naming is faster than and is able to interfere with shape naming (at least at first). Such a continuum suggests that speed of processing and interference effects are continuous variables which depend upon the degree of automatization of each task. This is supported by the following evidence.

Continuous nature of speed of processing. Numerous studies have shown that practice produces gradual, continuous increases in processing speed (e.g., Blackburn, 1936; Bryan & Harter, 1899; Logan, 1979; Shiffrin & Schneider, 1977) that follow a power law (Anderson, 1982; Kolers, 1976; Logan, 1988; Newell & Rosenbloom, 1980). MacLeod and Dunbar also examined this variable in their study. They continued to train subjects on the shape naming task with 144 trials/stimulus a day for 20 days. Reaction times showed gradual, progressive improvement with practice.

Continuous nature of interference effects. The pattern of interference effects observed in the MacLeod and Dunbar study also changed over the course of training on the shape naming task. As mentioned earlier, after 1 day of practice, there was no effect of shape names on color naming. After 5 days of training, however, shapes produced some interference, and after 20 days there was a large effect. That is, presenting a shape whose name conflicted with its ink color produced strong interference with the color naming response. The reverse pattern of results occurred for the shape naming task. After one

session of practice, conflicting ink color interfered with naming the shape. After 5 sessions, this interference was somewhat reduced, and after 20 sessions color no longer had an interfering effect on shape naming.

These data suggest that speed of processing and interference effects are continuous in nature, and that they are closely related to practice. Furthermore, they indicate that speed of processing and interference effects, alone, can not be used reliably to identify processes as controlled or automatic. These observations raise several important questions. What *is* the relationship between processes such as word reading, color naming and shape naming, and how do their interactions result in the pattern of effects observed? In particular, what kinds of mechanisms can account for continuous changes in both speed of processing and interference effects as a function of practice? Finally, and perhaps most importantly, how does attention relate to these phenomena?

The purpose of this paper is to provide a theoretical framework within which to address these questions. Using the principles of parallel distributed processing (PDP), we will describe a model of the Stroop effect in which both speed of processing and interference effects are related to a common, underlying variable that we call strength of processing. The model provides a mechanism for three attributes of automaticity. First, it shows how strength varies *continuously* as a function of practice; second, it shows how the *relative* strength of two competing processes determines the pattern of interference effects observed; and third, it shows how the strength of a process determines the extent to which it is governed by attention.

The model has direct implications for the standard method by which controlled and automatic processes are distinguished. It shows that two process that use qualitatively identical mechanisms and differ only in their strength, can exhibit differences in speed of processing and a pattern of interference effects that make it look as though one is automatic and the other is controlled. This suggests that these criteria — speed of processing, ability to produce interference, and susceptibility to interference — may be inadequate for distinguishing between controlled and automatic processing. This does not mean that the distinction between controlled and automatic processes is useless or invalid. Rather, the model shows that speed of processing differences and Stroop-like interference effects can emerge simply from differences in strength of processing, so that these phenomena may not provide a reliable basis for distinguishing controlled from automatic processes.

The Processing Framework

The information processing model we will describe was developed within the more general PDP framework described by Rumelhart, Hinton and McClelland (1986). Here, we outline some of the general characteristics of this framework. We then turn to the details of our implementation of a model of the Stroop effect.

Architectural characteristics. Processing within the PDP framework is assumed to take place in a system of connected modules. Each module consists of an ensemble of elementary processing units. Each unit is a simple information processing device that accumulates inputs from other units and adjusts its output continuously in response to these inputs.

Representation of information. Information is represented as a pattern of activation over the units in a module. The activation of each unit is a real valued number varying between a maximum and minimum value. Thus, information is represented in a graded fashion, and can accumulate and dissipate with time.

Processing. Processing occurs by the propagation of signals (spread of activation) from one module to another. This occurs via the connections that exist between the units in different modules. In general, there may be connections within as well as between modules, and connections may be bi-directional. However, for present purposes we adopt the simplification that there is a unidirectional flow of processing, starting at modules used to represent sensory input and proceeding "forward" or "bottom-up" to modules whose output governs the execution of overt responses.

Pathways and their strengths. A particular process is assumed to occur via a sequence of connected modules that form a *pathway*. Performance of a task requires that a processing pathway exist which allows the pattern of activation in the relevant sensory modules to generate — through propagation of activation across intermediate modules — an appropriate pattern of activation in the relevant output modules. The speed and accuracy with which a task is performed depends on the speed and accuracy with which information flows along the appropriate processing pathway. This, in turn, depends on the connections between the units that make up the modules in that pathway. We will demonstrate this in simulations shortly. We refer to this parameter as the *strength* of a pathway. Thus, the

speed and accuracy of performing a task depend on the strength of the pathway used in that task.

Interactions between processes. Individual modules can receive input from and send information to several other modules. As such, each can participate in several different processing pathways. Interactions between processes arise in this system when two different pathways rely on a common module — that is, when pathways intersect. If both processes are active, and the patterns of activation that each generates at their point of intersection are dissimilar, then *interference* will occur within that module, and processing will be impaired in one or both pathways. If the patterns of activation are very similar, this will lead to *facilitation.*

The intersection between two pathways can occur at any point in processing after the sensory stage. For example, interference at an intermediate stage is consistent with data reported by Shaffer (1975) and by Allport, Antonis and Reynolds (1972). Interference at the output stage would give rise to response competition, such as that observed in the Stroop task (cf. Dyer, 1973). The general view that interference effects arise whenever two processes rely on a common resource, or set of resources has been referred to as the multiple resources view (e.g., Allport, 1982; Hirst & Kalmar, 1987; Navon & Gopher, 1979; Wickens, 1984). Logan (1985) summarizes this position succinctly: "different tasks may depend on different resources, and dual-task interference occurs only when the tasks share common resources. Thus, the interference a particular task produces will not be an invariant characteristic of that task; rather, it will depend on the nature of the tasks it is combined with" (p.376). This point will be made explicit in the simulations we present below.

Attentional control. One way to avoid the interactions that occur at the intersection between two pathways is to modulate the information arriving along one of them. This is one of the primary functions of attention within this framework, and is consistent with the views on attention expressed by several other authors (Kahneman & Treisman, 1984; Logan, 1980; Treisman, 1960). In our system, modulation occurs by altering the responsiveness of the processing units in a pathway. In this way, attention can be used to control individual processes. However, this does not necessarily imply that attention requires a unique, or even distinct component of processing. As we shall see, attention can be thought of as an additional source of input which provides contextual support for the processing of signals within a selected pathway.

This framework can be used to account for many of the empirical phenomena associated with learning and automaticity. Schneider (1985) has used a similar approach to explain how performance in a category search task changes as a function of practice. Here, we focus on the significance that this approach has for selective attention, using the Stroop task as an example. In the next section we describe a simulation model of the Stroop task based on the processing principles discussed above. We then present a series of six simulations which demonstrate that this model is able to account for many of the empirical phenomena associated with automaticity, and for their gradual emergence as a function of practice. The first four simulations examine the attributes of automaticity evidenced in the Stroop task (viz., speed of processing and interference effects). The remaining simulations explore directly the relationship between processing and attention.

The Model

In this section, we describe the PDP mechanisms for processing, practice and attentional control that we used to simulate the Stroop task.

Architecture, Processing and the Representation of Information

The architecture of this model is depicted in Figure 1. The model consists of two processing pathways — one for processing color information, and the other for processing word information — both of which converge on a common response mechanism. Each pathway consists of a set of input units, a set of intermediate units, and a set of output units. Each of the input units in a given pathway projects to all of the intermediate units in that pathway. The intermediate units from both pathways project to all of the output units in the model. In addition, each unit is associated with a bias term, which is a constant value that is added to its net input (see below).

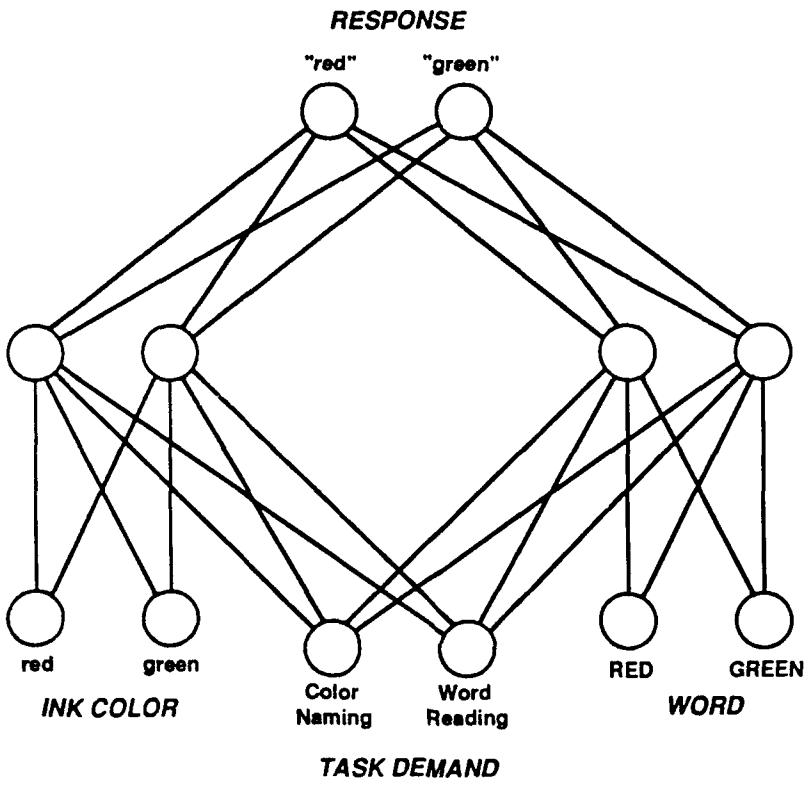


Figure 1. Network architecture. Units at the bottom are input units, and units at the top are the output (response) units.

Processing in this system is strictly feedforward. A stimulus is provided by activating units at the input level of the network. Activation then propagates to the intermediate units and, gradually, to the output units. A response occurs when sufficient activation has accumulated at one of the output units to exceed a response threshold. Reaction time is assumed to be linearly related to the number of processing cycles that it takes for this threshold to be exceeded (the response mechanism will be discussed in greater detail below). In addition to the units just described, there are also two task demand (or "attention") units — one for the color naming task and the other for the word reading task. These are connected to the intermediate units in the two processing pathways, and are used

to allocate attention to one or the other of them. Activation of a particular task demand unit sensitizes processing in the corresponding pathway, as will be explained shortly.

Individual stimuli and responses have discrete representations in this model. Each color is represented by a single input unit in the color pathway, and each word by a single input unit in the word pathway. Similarly, each output unit represents one potential response. We chose "local" representations of this kind to keep the model as simple and interpretable as possible. However, nothing in principle precludes the possibility that either inputs or outputs could be distributed over many units; and preliminary investigations indicate that our findings using local representations generalize to systems using distributed representations.

Mechanisms for Learning and the Time Course of Processing

The model is intended to provide an explanation of the relationship between learning and the time course of the psychological processes involved in the Stroop task. To date, PDP models that have addressed the time course of psychological processes have largely been distinct from those which address learning and memory. For example, McClelland (1979) presented a multilevel PDP system which provided an account of the time course of psychological processes, however this system did not include a learning algorithm. The back propagation algorithm described by Rumelhart, Hinton and Williams (1986) was introduced as a general learning mechanism which can be used in multilevel networks. However, PDP systems which have employed this algorithm generally have not simulated temporal phenomena such as reaction times. Here we describe each of these mechanisms and their limitations in greater detail. We then show how they can be brought together to provide a single system in which both learning and processing dynamics can be examined.

McClelland's (1979) cascade model provides a mechanism for simulating the time course of psychological processes. In this system, information is represented as the activation of units in a multilevel, feedforward network. Input is presented as a pattern of activation over units at the lowest level. Information gradually propagates upward, as units at each level update their activations based on the input they are receiving from lower levels. Eventually a pattern of activation develops over the units at the topmost level, where a response is generated. Units in this network update their activations based on a weighted

sum of the input they receive from units at the previous level in the network. Specifically, the net input at time (t) for $unit_j$ (at $level_n$) is calculated as:

$$net_j(t) = \sum_i a_i(t)w_{ij} \qquad \text{(Equation 1)}$$

where $a_i(t)$ is the activation of each $unit_i$ (at $level_{n-1}$) from which $unit_j$ receives input, and w_{ij} is the weight (or strength) of the connection from each $unit_i$ to $unit_j$. The activation of a unit is simply a running average of its net input over time:

$$a_j(t) = \overline{net}_j(t) = \tau\, net_j(t) + (1-\tau)\, \overline{net}_j(t-1) \qquad \text{(Equation 2)}$$

where $\overline{net}_j(t)$ is the time-average of the net input to $unit_j$, $net_j(t)$ is the net input to $unit_j$ at time (t), and τ is a rate constant. This time-averaging function is what establishes the time course of processing in this model. When τ is small the unit's activation will change slowly; with a larger τ it will change more quickly. One feature of Equation 2 is that, if the net input to a unit remains fixed, the unit's activation will approach an asymptotic value that is equal to this net input. As a result, McClelland demonstrates that with a constant input to the first layer in such a network, all of the units will approach an asymptotic activation value. Moreover, this value is determined strictly by the input to the network and the connections that exist between the units. Thus, given a particular input pattern, and sufficient time to settle, the network will always reach a stable state in which each unit has achieved a characteristic activation value.

One problem with the type of network used in the cascade model is that it is based on a linear activation function. That is, the activation of a unit is simply a weighted sum of the inputs it receives. It has been shown that networks which rely on linear update rules such as this, even if they are composed of multiple layers, suffer from fundamental computational limitations (*cf.* Rumelhart, Hinton, & McClelland, 1986 for a discussion). To overcome this problem, a network must have at least one layer of units between the input and output units that make use of a non-linear relation between input and output. Another problem with the cascade model, especially within the current context, is that it lacks any mechanism for learning. Both of these problems can be overcome if mechanisms are included that have been used in recent PDP models of learning.

The first step is to introduce non-linearity into processing. Typically, this has been done by using the logistic function to calculate the activation of a unit, based on its instantaneous net input:

$$a_j(t) = \text{logistic}(net_j(t)) = \frac{1}{1 + e^{-net_j(t)}} \qquad \text{(Equation 3)}$$

where $net_j(t)$ is given by Equation 1. The logistic function introduces non-linearity by constraining the activation of units to be between the values of 0 and 1 (see Figure 2). This non-linearity provides important behaviors, which we will discuss below (see "Attentional Selection"). However, as it stands, Equation 3 does not exhibit a gradual buildup of activation over time. The full response to a new input occurs in a single processing step at each level, and so the effects of a new input are propagated through the network in a single sweep through all of its levels. The dynamic properties of the cascade model can be introduced, however, if we assume — as the cascade model did — that the net input to a unit is averaged over time before the activation value is calculated. This gives us the *following activation rule:*

$$a_j(t) = \text{logistic}(\overline{net}_j(t)) \qquad \text{(Equation 4)}$$

where $\overline{net}_j(t)$ is defined as in Equation 2. The only difference between this activation rule and the one used in the cascade model is that the time-averaged net input to a unit is passed through the logistic function to arrive at its activation. We are still assured that the activation value will approach an asymptote which depends only on the input pattern and the connection strengths in the network. In fact, this asymptote is the same as the activation that the unit would assume without the use of time-averaging (to see this, consider the limiting case in which $\tau = 1$).

A number of learning rules have been described for single and multilevel networks using non-linear units. In the current model we used the generalized delta rule (also known as the back propagation learning algorithm) described by Rumelhart, Hinton and Williams (1986). Learning occurs by adjusting the connection strengths so as to reduce the difference between the output pattern produced by the network and the one desired in response to the current input. This difference is essentially a measure of the error in the performance of the network. Error reduction occurs by repeatedly cycling through the

following steps: (a) presenting an input pattern to be learned; (b) allowing the network to generate its asymptotic output pattern; (c) computing the difference between this output pattern and the one desired; (d) propagating information derived from this difference back to all of the intermediate units in the network; (e) allowing each unit to adjust its connection strengths based on this error information. By repeatedly applying this sequence of steps to each member of a set of input patterns, the network can be trained to approximate the desired output pattern for each input.

The non-linearity of the activation update rule discussed above is compatible with the back propagation algorithm, which only requires that the activation function be monotonic and continuous (i.e., differentiable). The logistic function satisfies this constraint. Furthermore, so long as units are allowed to reach their asymptotic activation values before error information is computed at the output level, then learning in this system is no different from systems which do not include a time-averaging component.

Variability and the Response Selection Mechanism

Processing variability. Even when human subjects appear to have mastered a task, they still exhibit variability in their response. This can be seen, for example, in the distribution of reaction times for a given task. In order to capture this variability, and to be able to model the variability of reaction time data, we introduce randomness into the model by adding normally distributed noise to the net input of each unit (except the input units).

Response mechanism. In addition to the variability in the activation process, the model also incorporates variability in the response mechanism. One successful way of modeling response variability has been to assume that the choice of a response is based on a random walk (Link, 1975) or a diffusion process (Ratcliff, 1978). In our adaptation of these ideas, we associate each possible response with an evidence accumulator which receives input from the output units of the network. At the beginning of each trial, all of the evidence accumulators are set to 0. In each time step of processing, each evidence accumulator adds a small amount of evidence to its accumulated total. The amount added is random and normally distributed, with mean μ based on the output of the network, and with fixed standard deviation σ. The mean μ is proportional to the difference between the activation of the corresponding unit and the activation of the most active alternative:

$$\mu_i = \alpha\,(act_i - max\_act_{j \neq i}) \qquad \text{(Equation 5)}$$

where α determines the rate of evidence accumulation. A response is generated when one of the accumulators reaches a fixed threshold. Throughout all of our simulations, the value of α was 0.1, the value of σ was 0.1, and the value of the threshold was 1.0.

This response selection mechanism may seem different from the rest of the network. For one thing, evidence is accumulated additively in the response selection mechanism, whereas running averages are used elsewhere in the network. For another, the response selection mechanism is linear, while the rest of the net is non-linear, and relies on this nonlinearity. In fact, it is easily shown that the additive diffusion process can be mimicked using linear running averages, by assuming that the response criterion gets smaller as processing goes on within a trial. The impact of introducing non-linearity into the evidence accumulator is less obvious. However, it need not exert a strong distorting effect, as long as the threshold is within the linear mid-portion of the accumulation function.

Attentional Selection

The role of attention in the model is to select one of two competing processes on the basis of the task instructions. In order for this to occur, one of two task demand specifications must be provided as input to the model: "respond to color" or "respond to word." We assume that this information is available as the output from some other module and results from encoding and interpreting the task instructions. Clearly, this is a highly flexible process, that can adapt to the wide variety of information processing tasks humans can perform. Our focus in this paper, however, is not on how task interpretation occurs or on how decisions concerning the allocation of attention are made. Rather, we are concerned with how information about the task and the corresponding allocation of attention influences processing in the pathways directly involved in performing the task itself. By focusing on the influences that attention has on processing, and specifying the mechanisms by which this occurs, we hope to show how attention interacts with strength of processing to determine the pattern of effects that are observed in the Stroop task.

Task information is represented in the model in the same way as any other information: as a pattern of activation over a set of processing units. For this purpose, two additional units

are included at the input level: one which represents the intention to name colors, and another for reading words. A particular task is specified by activating one of these "task demand" units. Task demand units modulate processing by adjusting the resting levels of units in the two main pathways, putting task appropriate units in the middle of their dynamic range and those for inappropriate units near the bottom where they will be relatively insensitive. We don't know whether, in actuality, attention is primarily excitatory (activating task-appropriate units), inhibitory (desensitizing inappropriate units) or — as we suspect — some of both. In any case, we assume that the connection strengths from the task demand units to intermediate units in each pathway are such that when the unit for a particular task is active, it sets the resting level of units in the appropriate pathway to the middle of their range, while units in the inappropriate pathway assume a more negative value. The modulatory influence that these changes in resting level have on processing is due to the non-linearity of the logistic activation function. To see how this occurs, let us examine this function in greater detail.

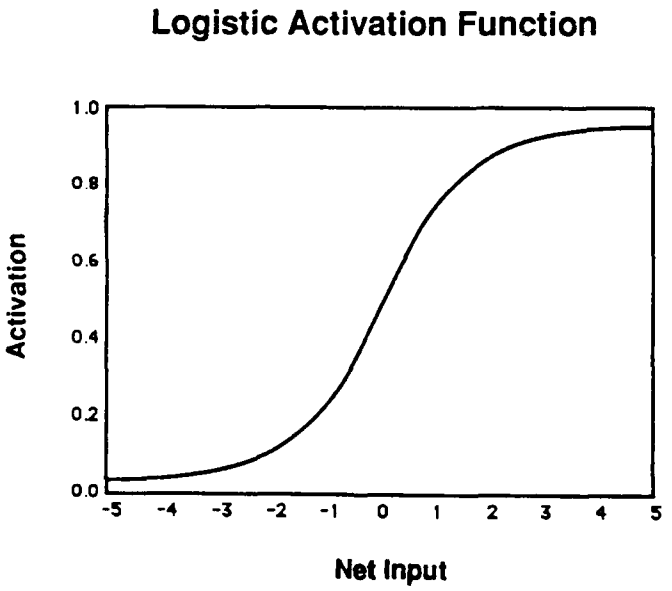


Figure 2. The logistic function. Note that the slope of this function is greatest when the net input is 0.0, and decreases when the net input is large in either the positive or negative directions.

As described by Equation 4, the activation of a unit is determined by the logistic of its net input. From Figure 2 it can be seen that the logistic function has roughly three regions. In the middle region — when the net input is close to zero — the relationship between net input and activation is more or less linear, with a slope close to 1. In this region, the activation of a unit is very responsive to changes in its net input. That is, changes in the net input will lead to significant changes in the unit's activation. In contrast, at each end of the logistic function the slope is dramatically reduced. In these regions — when the magnitude of the net input is large, either in a positive or a negative direction — changes in the input to a unit have a small effect on its activation. This feature was an important factor in our choice of a non-linear activation function, allowing the responsiveness of units to be modulated by adjusting their base levels of activation. This is accomplished by the activation of the task demand units.

In principle, task demand units are assumed to have connections to the intermediate units in each pathway, such that activation of a task demand unit drives the resting net input of units in the appropriate pathway toward zero, and units in competing pathways toward more negative values. Driving the net input of task-appropriate units toward zero places them in the most responsive region of their dynamic range, whereas making the net input of task-inappropriate units more negative places them in a "flatter" region of the activation function. In the current model, we implemented a simpler version of this general scheme. All intermediate units were assumed to have a negative bias, so that they were relatively insensitive at rest. Task demand units provided an amount of activation to intermediate units in the corresponding pathway that offset this negative bias, driving their net input to zero. Thus, task demand units had the effect of "sensitizing" units in the corresponding pathway, while units in the inappropriate pathway remained in a relatively insensitive state.

Finally, we should note that the connections between each task demand unit and all of the intermediate units within a given pathway are assumed to be uniform in strength, so that activation of a task demand unit does not, by itself, provide any information to a given pathway. Its effect is strictly modulatory.

Simulations

We implemented the mechanisms described above in a specific model of the Stroop task. In the following sections we describe how the model was used to simulate human performance in this task. We start by describing some of the general methods used in the simulations. We then describe four simulations which provide an explicit account of the attributes of automaticity and how they relate to practice. These are followed by two additional simulations which extend our consideration to issues concerning the relationship between attention and automaticity.

Simulation Methods

All simulations involved two phases: a training phase and a test phase.

Training phase. The network was trained to produce the correct response when information was presented in each of the two processing pathways. Training patterns were made up of a task specification and input to the corresponding pathway (see Table 1a). For example, an input pattern was "red-color-NULL," which activated the red input unit in the color pathway, the "respond to color" task demand unit, but did not activate any word input units. The network was trained to activate the red output unit as its response to this stimulus. Conflict and congruent stimuli were omitted from the training set, reflecting the assumption that, in ordinary experience, subjects rarely encounter these kinds of stimuli.

At the outset of training, the connection strengths between intermediate and output units were small random values. The connections between input units and intermediate units were assigned moderate values (+2 and -2) that generated a distinct representation of each input at the intermediate level. This set of strengths reflects the assumption that, early in their experience, subjects are able to successfully encode sensory information (e.g., colors and word forms) at an intermediate level of representation, but are unable to map these onto appropriate verbal responses. This ability only comes with training. This initial state of the network also allowed us to capture the power law associated with training, that we will discuss below (see Simulation 3).

Table 1a. Training Stimuli.

	Task demand	*Color input*	*Word input*	*Output*
a)	color	red	——	"red"
b)	color	green	——	"green"
c)	word	——	RED	"red"
d)	word	——	GREEN	"green"

Table 1b. Test Stimuli.*

Stimulus Type	*Task demand*	*Color input*	*Word input*
Color naming:			
task specification	color	——	——
control	color	red	——
conflict	color	red	GREEN
congruent	color	red	RED
Word reading:			
task specification	word	——	——
control	word	——	RED
conflict	word	green	RED
congruent	word	red	RED

* Only those stimuli for which "red" was the correct response are shown. The network was also tested with the corresponding stimuli for which "green" was the correct response.

The influence of attention was implemented in the simplest way possible. Bias parameters for intermediate units and connection strengths from the task demand units were chosen so that when a particular task demand unit was on, the intermediate units in the attended

pathway had a base net input of 0.0, and were thus maximally responsive to input (see above). Units in the unattended pathway had a much lower base activation. The value of the base activation of units in the unattended pathway (determined by their negative bias) reflected the effectiveness of filtering in a given task, and was allowed to vary from experiment to experiment (see below).

In each training trial, an input pattern was presented to the network, and all of the units were allowed to reach their asymptotic values.[1] Difference terms were then computed by comparing the actual activation with the desired activation value for each output unit. These difference terms were treated as error signals that were then used to calculate changes to the connection strengths following the back propagation learning procedure (Rumelhart, Hinton & Williams, 1986).[2] All of the connections along the word and color processing pathways were modifiable, and their values were set using the learning procedure just described. However, the connections from the task demand units to the intermediate units in each pathway, and the bias terms that established the resting activations of these units were assumed to be unmodifiable. Training proceeded until the network was capable of correctly processing all of the test stimuli (see below: "Test Phase").

One purpose of the model is to account for the relationship between practice effects and automaticity. In the context of the Stroop task, it has been proposed that word reading is more highly practiced than color naming (Brown, 1915; MacLeod & Dunbar, 1988; Posner & Snyder, 1975). In order to model this difference in practice, we gave the network differential amounts of training on the word and color patterns. Every word pattern was presented in every epoch, while the probability of a color pattern being presented in a given epoch was 0.1. Thus, on average, word patterns were seen ten times as often as color patterns, and at any given point during training, the network had received a greater amount of practice with word reading than color naming.

---

[1] Processing was deterministic during training; that is, units were not subject to noise. Individual simulations using noise during training indicated that this did not significantly alter the results, and the elimination of noise in this phase substantially reduced the length and number of simulations required to arrive at a normative set of results.

[2] Connection strengths were updated after each sweep through the set of training patterns. Learning rate was 0.1 and momentum was 0.0.

Figure 3 displays the strengths on all of the connections in the network at the end of training. As expected, they were stronger in the word pathway than in the color pathway, due to the greater frequency of word training.[1]

Test phase. The network was tested on the 12 input patterns corresponding to all possible stimuli in a Stroop task in which there are two possible responses (e.g., "red" and "green"). These patterns represented the control stimulus, the congruent stimulus, and the conflict stimulus for each of the two inputs (red or green) in each of the two tasks (word reading and color naming) (see Table 1b). Presentation of a particular pattern consisted of activating the appropriate input unit(s) and task demand unit. For example, one of the conflict stimuli in the color naming task (the word GREEN in red ink) was presented by activating the red color input unit, the "attend to color" task demand unit, and the GREEN word input unit.

---

[1] We focus on frequency of training as the primary difference between word reading and color naming because this has been the emphasis in the literature. However, other differences between these tasks might also be important. For example, it seems likely that word reading is also a more consistently mapped task than color naming: a particular sequence of letters is almost invariably associated with the word they represent (even if the word itself has an ambiguous meaning); however colors are often associated with words other than their name (e.g., red is associated with heat, embarrassment and "stop"). While this point has not been emphasized with regard to the Stroop task, it is a well established finding that consistent mapping leads to the development of automaticity, while variable mapping impedes it (e.g., Logan, 1979; Shiffrin & Schneider, 1977). Our model captures this fact: the more consistently a stimulus is related to a particular response, the stronger will be the connections for processing that stimulus. Although in this paper we will focus on frequency (i.e., *amount* of practice) as a determinant of pathway strength, it should be kept in mind that consistency of practice is an equally important variable that may be a significant factor underlying the Stroop effect.

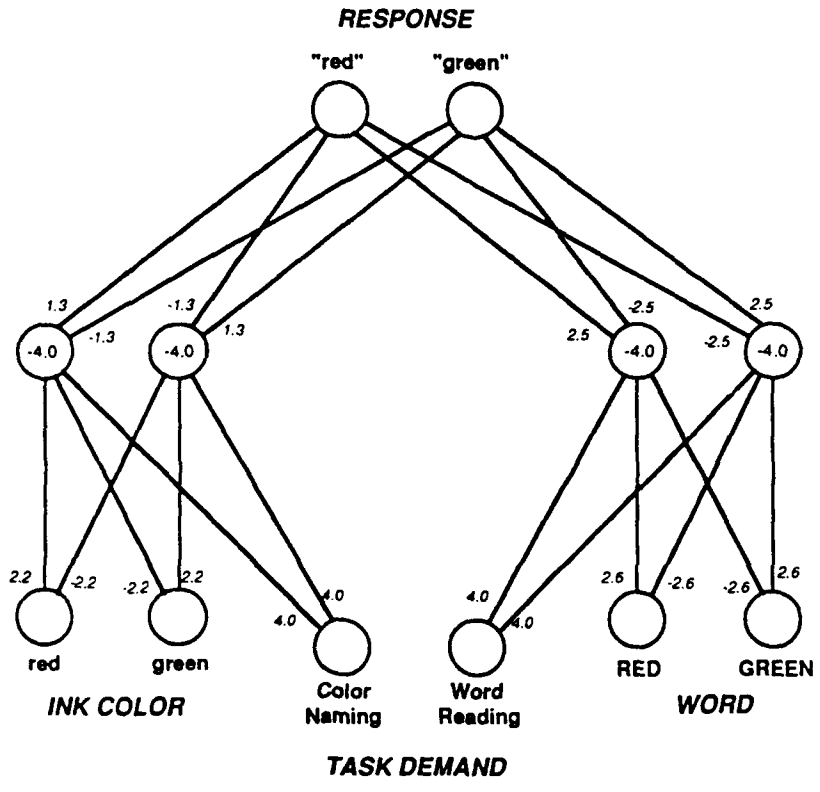


Figure 3. Diagram of the network showing the connection strengths after training on the word reading and color naming tasks. Strengths are shown next to connections; biases on the intermediate units are shown inside the units. Attention strengths (i.e., from task demand units to intermediate units) were fixed, as were biases for the intermediate units. The values were chosen so that when the task demand unit was on, the base input for units in the corresponding pathway was 0.0, while the base input to units in the other pathway was in range of -4.0 to -4.9, depending upon the experiment (see text).

Each test trial began by activating the appropriate task demand unit, and allowing the activation of all units to reach asymptote. This put the network in a "ready" state corresponding to the appropriate task. At this point, the intermediate units in the selected pathway and all of the output units had resting activation levels of 0.5, while intermediate units in the competing pathway were relatively inactive (activations of approximately 0.01). The test pattern was then presented, and the system was allowed to cycle until the activation accumulated from one of the output units exceeded the response threshold. A value of 1.0

was used for the response threshold in all simulations. The number of cycles required to exceed this threshold was recorded as the "reaction time" to that input. The system was then reset, and the next trial began. Data values reported below represent the mean value of 100 trials run for each condition. A representative sample of the reaction time distributions obtained in this way is given in Figure 4. This shows the skewed distribution typical of human data and standard random walk models (e.g., Ratcliff, 1978).

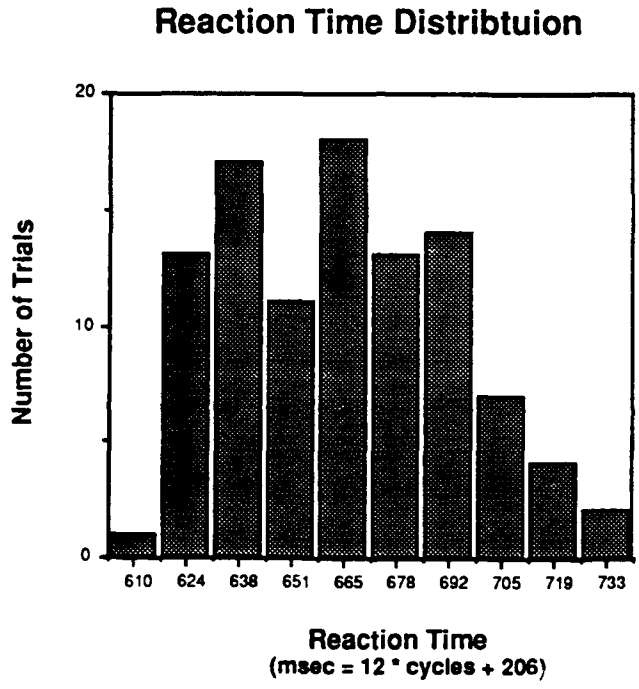


Figure 4. Distribution of reaction times for 100 trials of color naming (control condition) from Simulation 1.

To simplify comparison between empirical reaction times and the model's performance, we report simulation reaction times as transformed values. For each simulation, we performed a linear regression of the simulation data on the empirical data. Simulation data are reported

here as the number of cycles transformed by this regression equation.[1] Regression equations are provided in the figure(s) accompanying each simulation.

Free parameters. We have undertaken a large number of simulation experiments, varying different parameters of the model and examining how these affected the model's ability to account for the basic form of the empirical phenomena. Appendix A describes the parameter values used in the reported simulations, as well as several trade-offs and interactions between parameters that we encountered. In general, we strove to use a single set of parameters for all simulations. However, in comparing the results of different empirical studies it became apparent that nominally identical experimental conditions sometimes produce rather different interference and facilitation effects. In particular, in experiments where subjects have to say the color of the ink in which words are actually written, interference effects may be more than twice as large as in experiments where color and word information occur in physically different locations. It seemed likely that this difference reflected differences in subject's ability to selectively modulate processing of task-relevant and task-irrelevant information. To capture this, we allowed the strength of the attentional effect to be adjusted separately for each simulation. This was done by varying the resting activation level of units in the unattended channel, thereby placing them in a more or less responsive state.

Strength of Processing

Simulation 1. The Basic Stroop Effect

The purpose of the first simulation was to provide an account for the set of empirical findings that comprise the basic Stroop effect. These are displayed in Figure 5a, and are described below.

---

[1] In all cases, the intercept of the regression equation was positive, reflecting components of processing (e.g. early visual processing and response execution) not simulated by the model. The intercept value for all of simulations was in the range of 200-500 msecs.

Word reading is faster than color naming. The time to read a color word is about 350-450 msec, whereas the time to name a color patch or a row of colored X's is 550-650 msec. Thus word reading is about 200 msec faster than color naming (cf. Cattell, 1886; Dyer, 1973; Glaser & Glaser, 1982).

Word reading is not affected by ink color. Ink color has virtually no effect on the amount of time that it takes to read the word. That is, reaction times to read the word in the conflict and congruent conditions are the same as in the control condition. This phenomenon was originally discovered by Stroop (1935) and can be seen in the flat shape of the graph for word reading in Figure 5a. This finding is extremely robust and is very difficult to disrupt. Even when the ink color appears before the word it does not interfere with word reading (Glaser & Glaser, 1982). It is only when the task is changed radically that the ink color will interfere with word reading (Dunbar & MacLeod, 1984; Gumenik & Glass, 1970).

Words can influence color naming. A conflicting word produces a substantial increase in reaction time for naming the ink color relative to the control condition. The amount of interference is variable, but is usually around 100 msec (e.g., Dunbar & MacLeod, 1984; Glaser & Glaser, 1982; Kahneman & Chajczyk, 1983). This finding is also extremely robust and nearly all subjects show the effect. Even when the word and the ink color are presented in different spatial locations (e.g., the word is placed above a color patch) the word still interferes with ink color naming (Gatti & Egeth, 1978; Kahneman & Henik, 1981). In the congruent condition the word facilitates ink naming, producing a decrease in reaction time relative to the control condition (Hintzman et al., 1972). The amount of facilitation can range from about 20 msec (Regan, 1978) to about 50 msec (Kahneman & Chajczyk, 1983).

There is less facilitation than interference. Congruent stimuli have not been used as extensively as conflict stimuli, but the general finding is that the amount of facilitation obtained is much less than the amount of interference (Dunbar & MacLeod, 1984).

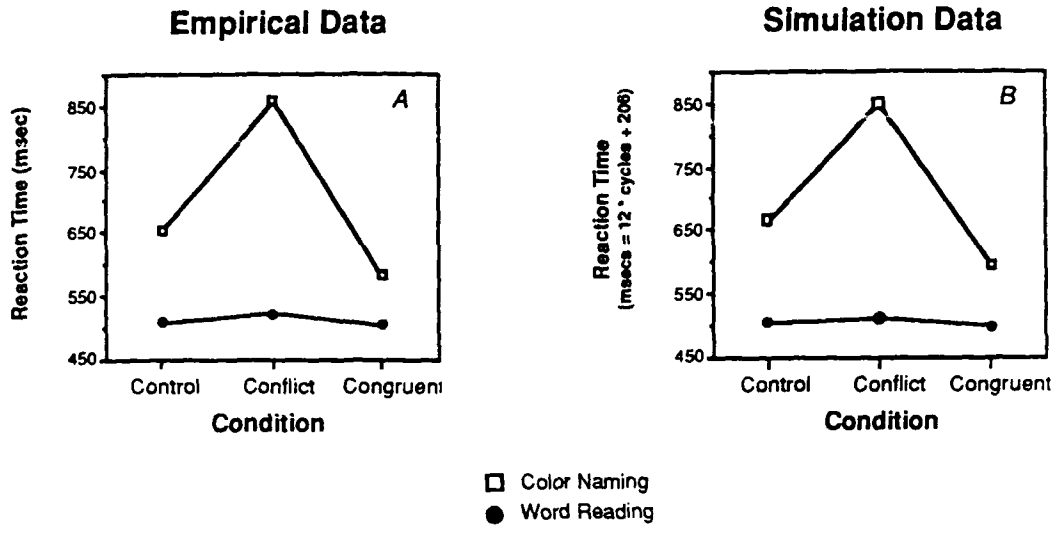


Figure 5. Performance data for the standard Stroop task. Panel A shows data from an empirical study (after Dunbar & MacLeod, 1984). Panel B shows the results of the model's simulation of this data.

Figure 5a shows the findings in a standard Stroop experiment (Dunbar & MacLeod, 1984). Figure 5b presents the results of our simulation, which reproduces all of the empirical effects. These are explained as follows.

Word reading was faster than color naming in the simulation because differential amounts of training led to the development of a stronger pathway for the processing of word information than color information. The fact that the network was trained more extensively with word stimuli than with colors meant that units in the word pathway had a greater number of trials in which to increment their connection strengths (see Figure 3). Stronger connections resulted in larger changes to the net input — and therefore to the activation — of word units in each processing cycle (see Equations 1 and 2). This allowed activation to accumulate at the output level more rapidly in the word pathway than the color naming pathway. The faster the correct response unit accumulates activation (and competing units become inhibited), the faster the response threshold will be exceeded. Thus, the strength of a pathway determines its speed of processing.

The difference in the strength of the two pathways also explains the difference in interference effects between the two tasks. First, consider the failure of color information to affect the word reading task. Here, activation of the task demand unit puts intermediate units in the word reading pathway in a responsive state, so that information flows effectively along this pathway. In contrast, because no attention is allocated to the color pathway, units in this pathway remain in an unresponsive state, and accumulation of information at the level of the intermediate units is severely attenuated. Furthermore, because the connections from intermediate to output units are weaker in the color pathway, what information does accumulate on intermediate units is transmitted to the output level more weakly than information flowing along the word pathway. Both of these factors diminish the impact that color information has on the network's response to a word. As such, reaction time in the word reading task is only very slightly affected by the presence of either congruent or conflicting color input.

Very different results occur when color naming is the task. Attention is now allocated to this pathway, so that the intermediate units are placed in a responsive part of their dynamic range, and information flows unattenuated to the output level. It is now the units in the word pathway that are relatively unresponsive. However, because of the stronger connections in the word pathway, more activation can build up at the intermediate unit level. The amount of this accumulation is greater than it was for color units in the word reading task.[1] Furthermore, the connections from the intermediate to output units in this pathway are also stronger than in the color pathway, so that what information accumulates on the intermediate units has a greater influence at the output level. Thus, some information flows along the word pathway even in the absence of the allocation of

---

[1] As an example, consider the case in which the RED word input unit is activated. This has an excitatory connection to the leftmost intermediate unit in the word pathway, with a strength of 2.63. In the absence of input from the task demand unit (and ignoring the effects of noise), this intermediate unit receives a net input of 2.63 + (-4_{bias}) = -1.37. After passing this through the logistic activation function, we arrive at an asymptotic activation of .2 for this unit. This will be the amount contributed to the net input of the "red" output unit. Now consider the situation for the color naming pathway. There, the strength of the connection from the red input unit to the corresponding intermediate unit is only 2.20. In the absence of task demand activation, the intermediate unit will have a net input of 2.20 + (-4_{bias}) = -1.8 which, when passed through the logistic function, results in an activation of 0.14. Thus, in the absence of attention, activation of an intermediate color unit is lower than that of a corresponding word pathway unit.

attention. Although this flow of information is only partial, and is not sufficient to determine which response is made, it is enough to affect the speed with which a response is made, thus producing interference in the color naming task. This processing of information in the word pathway without the allocation of attention captures the "involuntariness" of word reading, and accounts for the interference and facilitation effects that are observed. All of these effects are attributable to the fact that the word reading pathway is stronger (i.e., has stronger connections) than the color naming pathway.

The fourth finding is that the amount of interference is consistently larger than the amount of facilitation. In the model there are two factors that contribute to this result. One is the non-linearity of the activation function. This imposes a ceiling on the activation of the correct response unit, which leads to an asymmetry between the effects of the excitation it receives from the irrelevant pathway in the congruent condition, and the inhibition it receives in the conflict condition. To see this more clearly, consider the idealized situation depicted in Figure 6a.

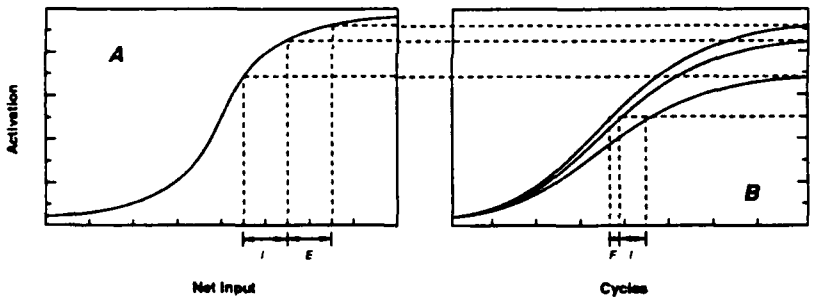


Figure 6. Mechanisms underlying the asymmetry between interference and facilitation effects. Panel A shows the effects of equal amounts of excitation (E) and inhibition (I) from a competing pathway on the asymptotic activation of an output unit. Panel B shows the effects of these different asymptotic levels of activation on the time to reach a particular level of activation (F = facilitation; I = interference).

In this figure, the activation function for the correct response unit is shown. Its asymptotic activation is plotted for each of the three experimental conditions in a color naming trial. Note that activation is highest in the congruent condition and lowest in the conflict condition. This is because in the congruent condition, the irrelevant pathway contributes

excitatory input to the response unit, increasing its net input; whereas in the conflict condition it contributes inhibition, decreasing the response unit's net input. Note that, although the increase in net input in the congruent condition is equal in magnitude to the decrease in the conflict condition, the effect on the activation of the response unit is not symmetric: inhibition has a greater effect than excitation. This is because this unit is in a non-linear region of the logistic activation function. In this region, increasing the net input has less of an effect on activation than decreasing it.[1]

Figure 6a shows the asymptotic activation values for the response unit in each of the three conditions. Figure 6b plots the rise in response unit activation, over time, toward each of these asymptotic values. Note that, at any given point in time, the difference in activation between the control and conflict conditions is greater than the difference between the control and congruent conditions. This shows that, throughout the course of processing, inhibition has a greater influence than excitation on the accumulation of evidence at the output level. Thus, the non-linearity of the logistic function, and its interaction with the dynamics of processing help to produce the asymmetry between the size of interference and facilitation effects observed in the simulation.

A second factor also contributes to the asymmetry in the magnitudes of interference and facilitation. This is the basically negatively accelerating form of the curve relating activation to cycles of processing. This negatively accelerating curve is an inherent property of the cascade mechanism (time averaging of net inputs), and would tend to cause a slight asymmetry in the interference and facilitation effects even if interference and facilitation had exactly equal and opposite effects on asymptotic activation. However, this is a relatively weak effect, and is not sufficient in and of itself to account for the greater than 2:1 ratio of interference to facilitation that is typically observed.

---

[1] The reason that output activations fall in this region has to do with the nature of the activation function and training in this system. Early in training, the connections to an output unit are small, so that the net input it receives — regardless of the input pattern being presented — is close to 0.0, and its activation is close to 0.5. If the correct response to a particular input pattern requires that output unit to have an activation value of 1.0, then learning will progressively adjust its connections so that its activation shifts from 0.5 to a value closer to 1.0 when that input pattern is present. The region between 0.5 and 1.0 (for units whose output should be 1.0) is precisely the region of the logistic function that produces the asymmetry between interference and facilitation observed in our simulations.

Neither the logistic function nor the cascade mechanism was included in the model specifically to produce an asymmetry between interference and facilitation. The logistic function was included in order to introduce non-linearity into processing for the purpose of computational generality (see above: "The Mechanisms Underlying Learning and the Time Course of Processing"), and to allow attention to modulate the responsiveness of units in the processing pathways. The cascade mechanism was introduced in order to model the dynamics of processing. The fact that these mechanisms led to an asymmetry between interference and facilitation is a by-product of these computationally motivated features of the model.

It is worth noting that most theories have been unable to account for this asymmetry in terms of a single processing mechanism. In fact, several authors have argued that separate processing mechanisms are responsible for interference and facilitation effects (e.g., Glaser & Glaser, 1982; MacLeod & Dunbar, 1988). Although this remains a logical possibility, our model demonstrates that this is not necessarily the case. We believe that the failure of previous theories to account for this asymmetry in terms of a single mechanism has been due to their reliance, either explicitly or implicitly, on linear processing mechanisms.

Simulation 2. Stimulus Onset Asynchrony Effects: Speed of Processing and Pathway Strength

The results of the previous simulation demonstrate that the strength of a pathway determines both speed of processing, and whether or not one process will influence (interfere with or facilitate) another. In this simulation we demonstrate that pathway strength — and not just speed of processing — is responsible for interference and facilitation effects.

The speed of processing account of the Stroop effect assumes that the faster finishing time of the word reading process is responsible for the asymmetry in interference effects between word reading and color naming. If no other factors are assumed, then this account

predicts that the Stroop effect can be reversed by presenting color information before the word.[1]

Glaser and Glaser (1982) tested this prediction and found no support for it: color information failed to interfere with word reading even when color information preceded the word by 400 msec. Indeed, they found no effect of colors on words over stimulus onset asynchronies ranging from -400 msec (color preceding word) to 400 msec (word preceding color). This is shown in the lower part of Figure 7a, which presents data from the word reading condition of one of their experiments.

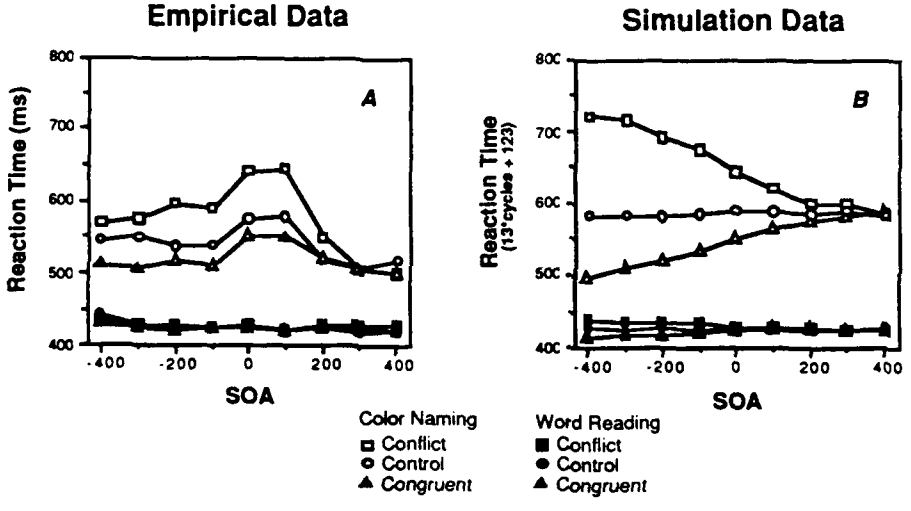


Figure 7. Effects of varying SOA between word and color stimuli in the color naming and word reading tasks. Panel A shows data from an empirical study (after Glaser & Glaser, 1982). Panel B shows the results of the model's simulation of these effects.

We simulated the Glaser and Glaser experiment by activating the color input unit before and after the word input unit. This was done at a number of cycles corresponding to the SOA's

[1] This requires spatial separation of color and word stimuli. This reduces, but does not eliminate the standard set of effects (see Gatti & Egeth, 1978).

used in the actual experiment.[1] In order to simulate the reduced interference and facilitation effects observed at the 0 msec SOA in this experiment — in comparison with the standard experiment using integral stimuli — we increased the size of the attentional effect for both pathways by decreasing the resting net input to units in the unattended from -4.0 to -4.9. The results of this simulation are presented in Figure 7b.

The model shows little interference of color on word, regardless of SOA, just as is seen in Glaser and Glaser's data. It is true that when color precedes word, the model shows a slight effect of color on word, but it is much smaller than the effect of word on color (the maximum, and what appears to be the asymptotic amount of interference produced by colors on words is substantially less than the amount of interference produced by words on colors at the 0 msec SOA). In this way, the model concurs with the empirical data, suggesting that differential speed of processing is not the sole source of interference observed in the Stroop task. The model shows that interference is substantially influenced by differences in strength of processing: when attention is withdrawn from the weaker pathway, it is able to produce less activation at the output level than the stronger pathway is able to produce when attention is withdrawn from it. As a result, weaker pathways produce less interference, independent of their finishing time.

Nevertheless, there is a discrepancy between the model and the empirical data in Figure 7. The simulation shows *some* influence of color on word reading when the color is presented sufficiently in advance of the word, whereas the subjects do not. In fact, Phaff (1986) has reported empirical data of Neumann's (1980) which indicate that, under some conditions, early-appearing colors *can* produce a small amount of interference with word reading, just as the model leads us to expect. It is unclear, therefore, whether this mismatch between the simulation and the Glaser and Glaser data represents a limitation of the model, or the involvement — in their experiment — of additional processes that are not central to the

---

[1] The number of cycles corresponding to each SOA was determined in the following manner. The simulation was tested at the 0 msec SOA (color and word presented simultaneously, as in Simulation 1). A regression was performed of these data on the Glaser and Glaser data at the 0 msec SOA. Other SOAs were then divided by the regression coefficient to arrive at the number of cycles to be used for each SOA in the simulation.

Stroop effect. The latter possibility is suggested by another discrepancy between our simulation and the empirical results.

In Glaser and Glaser's experiment, subjects showed very little interference in color naming when the word appeared more than 200 msec in advance of the color (see upper part of Figure 7a). In their original analysis this was attributed to strategic effects. More recently, they have suggested that a process of habituation may be involved (W. Glaser, personal communication). Our model does not include such a process, and this may be why the simulation shows greater rather than lesser amounts of interference at the longer negative SOAs. Note, however, that if habituation applies to color stimuli as it does to words, then it would also tend to reduce any effect that colors have on word reading at the longer SOAs. If this effect were small to start with, it might be entirely eliminated by habituation. This may explain why Glaser and Glaser failed to observe any effect of colors on words at long SOAs but, owing to lack of a habituation process in our model, this effect was observed in the simulation.

In summary, although the model does not capture all aspects of the empirical data, it clearly demonstrates our central point: that differential strength of processing can explain why presenting a weaker stimulus before a stronger one fails to compensate for differences in processing speed with regard to interference and facilitation effects.

Practice Effects

A primary purpose of this model is to show how the changes in strength that occur with practice can lead to the kinds of changes in speed of processing and interference effects observed for human subjects. These phenomena are addressed by the following two simulations.

Simulation 3. The Power Law

Numerous studies have demonstrated that the increases in speed of processing that occur with practice follow a power law (Anderson, 1982; Kolers, 1976; Logan, 1988; Newell & Rosenbloom, 1980). This finding is so common that some authors have suggested that, in order to be taken seriously, any model of automaticity must demonstrate this behavior (e.g., Logan, 1988). The power law for reaction time (RT) as a function of number of training trials (N) has the following form:

$$RT = a + bN^{-c} \qquad \text{(Equation 5)}$$

where a is the asymptotic value of the reaction time, b is the difference between initial and asymptotic performance, and c is the learning rate associated with the process. When this function is plotted in log-log coordinates, reaction time should appear as a linear function of number of trials, with slope c. Typically, RT is the mean of the distribution of reaction times for a process at a given point in training. Recently, Logan (1988) has shown that, at least for some tasks, the standard deviation of this distribution also decreases with training according to a power law, and that this occurs at the *same rate* as the decrease in mean reaction time (i.e., the coefficient c is the same for both functions). This means that, in log-log coordinates, the plot of reaction times should be parallel to the plot of standard deviations.

In order to assess the current model for these properties, we trained the network on the color naming task for 100,000 epochs. At regular intervals, the network was given 100 test trials (control condition) on this task. Figure 8 shows the log of the mean reaction time minus its estimated asymptote and the log of the standard deviation minus its estimated asymptote, each plotted against the log of the number of training trials. Lines represent the best fit to these data using a logarithmic regression; equations for each regression are also shown, along with squared correlation between observed and predicted values. Both mean reaction time and standard deviation are closely approximated by power functions of training. Furthermore, the exponents of the two functions are very similar, and are within the range of variation exhibited by Logan's empirical data.

Learning follows a power law for two reasons. First, learning in the network is error-driven. That is, the amount that each connection weight is changed is based on how much each output unit activation differs from its desired (target) value. Early in training this difference is likely to be large (otherwise the problem would already be solved), so large changes will be made to the connection strengths. As the appropriate set of strengths develops, the error will get smaller and so too will the changes made to the connections in each training trial. We should note, however, that although weight changes will get smaller with practice, they will continue to occur as long as there is training. This is because target values are taken to be 1.0 for units which should be active, and 0.0 for all others. These targets values can never actually be reached with finite input to units using the logistic

activation function (see Figure 2). Thus there is always some "error" and therefore always some additional strengthening of connections that is possible. However, this strengthening will get progressively less with training, and therefore improvements in reaction time will become less as well.

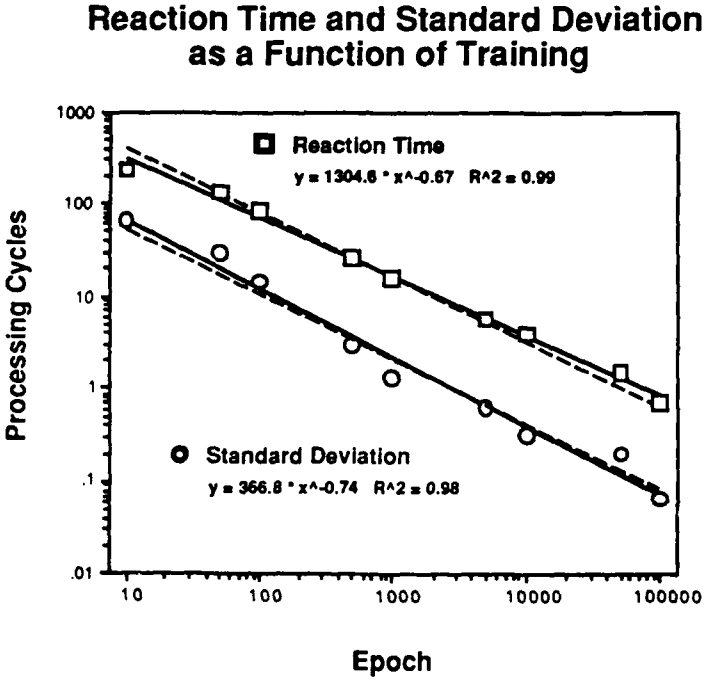


Figure 8. Log-log plot of the mean and standard deviation for reaction time at various points during training on the color naming task. Regression equations are for mean reaction times and standard deviations separately, and are plotted as solid lines. Dashed lines show the regression which best fits both sets of data simultaneously.

A second reason for the deceleration of improvements in reaction time with practice is that as connections get stronger, subsequent increases in strength have less of an influence on activation (and therefore reaction time). This is due to the non-linearity of the activation function: once a connection (or set of connections) is strong enough to produce an activation close to 0.0 or 1.0, further changes will have little effect on that unit. Thus, smaller changes in strength compounded with the smaller effects that such changes have combine to produce the pattern of doubly diminishing returns that is captured by the log-log relation between reaction time and practice.

The arguments just provided apply only when representations in the next to last layer have already been fairly well established, and training involves primarily the connections between this layer and the output layer. In training a multilayer network from scratch using back propagation, there is a long initial phase of slow learning, followed by one or more periods of rapid acceleration, and then finally a phase which follows a power law. Accordingly, when both the input and output layers of connections in our network had to be learned, improvements in reaction time did not follow a power law from the start of training. Adherence to the power law occurred only when meaningful and moderately strong connections from the input units to the intermediate intermediate units were already in place at the beginning of training. Although these input connections were modifiable — and were augmented during training — their initial values had to be such that the network could do the task by modifying only the output connections.

While some might take these findings as an indictment of the back propagation learning algorithm, we suggest that they may reflect constraints on the applicability of the power law: it may apply to only certain types of learning. Specifically, it may not apply to situations in which an intermediate representation must be constructed to perform a task. These may involve more than one phase of learning, as is observed in back propagation networks when more than one layer of weights must be learned. Along these lines, we have used a back propagation model to capture the stage-like character of learning reported in a set of developmental tasks (McClelland, in press), and Schneider and Oliver (in press) have begun to explore how back propagation nets can capture multiphase learning observed for certain tasks in adults.

Simulation 4. Practice Effects and the Development of Automaticity

Having demonstrated that the current model conforms to standardly accepted laws of learning, we now apply it to empirical data concerning learning, and the effect that learning has on interference effects. MacLeod and Dunbar (1988) have shown that both speed of processing *and* the ability of one process to interfere with (or facilitate) another are affected by the relative amounts of training that subjects have received on each. In their experiments, subjects were taught to associate a different color name to each of four different shapes. During the training phase, the shapes were all presented in a neutral color (white) and subjects practiced naming these for 20 days. Mean reaction times were calculated for each day of training (these are shown below in Figure 9). At the conclusion

of 1, 5, and 20 days of practice, subjects were tested with neutral, conflict, and congruent stimuli in both the shape naming and color naming tasks.[1] The results of this experiment can be summarized as follows (these also appear in Figure 12a, below):

- After 1 day (72 trials/stimulus) of practice on shape naming, this was still more than 100 msec slower than color naming. The shapes had no effect on the time to name the ink colors. However, the ink colors produced both interference and facilitation in the shape naming task. The amount of interference was greater than the amount of facilitation.

- After 5 days (504 trials/stimulus) of practice, shape naming was significantly faster than on day 1. In addition, the shapes now interfered with color naming, although they did not produce facilitation. The colors continued to produce both interference and facilitation in shape naming.

- After 20 days (2,520 trials/stimulus) of practice, shape naming was slightly faster than ink naming. The shapes produced a large amount of interference and a small amount of facilitation in naming colors. The colors now produced much smaller amounts of facilitation and interference in shape naming.

MacLeod and Dunbar argued that these data contradict the idea that the attributes of automaticity are all or none. They suggested instead that a continuum of automaticity exists, in which it is the *relative* amount of training on two tasks that determines the nature of the interactions between them. The current model provides a mechanisms for this.

To simulate the MacLeod and Dunbar experiments, we used the network from the previous simulations (which had already been trained on color naming and word reading), adding a new pathway that was used for shape naming. This pathway was identical in all respects to the two pre-existing ones, except that it had not received any training (see Figure 9). As with the color and word pathways, it was given a set of initial connection strengths from

---

[1] For the shape naming task these were: (a) shapes in a neutral color (control condition), (b) shapes in a color that was inconsistent with the shape name (conflict condition) and, (c) shapes in a color that was consistent with the shape name (congruent condition). The same stimuli were used for the color naming task, except that the control condition used a neutral shape (square).

the input to the intermediate units that allowed it to generate a useful representation at the level of the intermediate units, while small random strengths were assigned to the connections between the intermediate and output units.

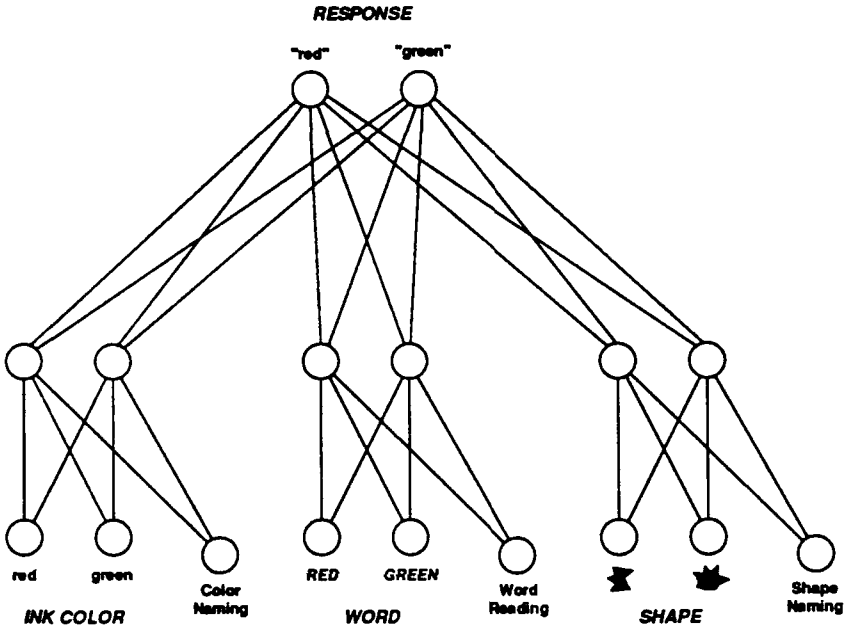


Figure 9. The network architecture used to simulate the shape naming experiments conducted by MacLeod and Dunbar (1988).

Using this expanded network, we examined how practice on a novel task (shape naming) affects its interaction with another task that has already received a moderate amount of training (color naming). The word pathway was not used in this simulation.

The simulation involved a series of alternating phases of training and testing, as in the empirical study. During each training phase, the network was presented with only the two control shape patterns. Both these patterns were presented in every training epoch. Although only two shape stimuli were used in the simulation, the network received exactly the same number of training exposures per stimulus that subjects received in the experiment. During testing, the network was presented with the conflict and congruent as

well as the control stimuli in each task condition (color naming and shape naming). Reaction times to each stimulus type were recorded and averaged over each condition.

In the empirical study, test sessions gave subjects additional practice with the stimulus items, including conflict and congruent stimuli. In order to accurately simulate these circumstances, we allowed the network to continue to adjust its connection strengths (in both the color and shape pathways) during each test phase. The network received exactly the same number of exposures to each test stimulus as did human subjects. Furthermore, these were blocked by task and presented in the same sequence as in the empirical study. The network was tested after it had received the same number of training exposures per stimulus received by subjects on days 1 (72), 5 (504) and 20 (2,520).

We simulated two components of the MacLeod and Dunbar data: 1) changes in the speed of shape naming with practice, and 2) changes in interference effects between shape naming and color naming. We consider each of these in turn.

Practice effects. According to the findings for other tasks in which practice leads to automaticity, improvements in reaction time for shape naming should have followed a power law. Mean reaction time for shape naming on each day of training are shown in Figure 10a (solid squares). These data are reasonably well fit by a power function (see Figure 10b). Figure 10a also shows the performance of the model as we have described it so far (open triangles). The network exhibited significantly longer reaction times than subjects early in training. This suggested that, early on, subjects might be performing the task in a different way than they did later. This agrees with the general idea that flexible, general purpose resources are required to perform novel tasks, and only with practice do automatic mechanisms come into play. Strategic (e.g., Posner & Snyder, 1975), controlled (Shiffrin & Schneider, 1977) and algorithm-based (Logan, 1988) processes would all fit into this category. The present model was not intended to address the mechanisms underlying such processes in detail. However, in order to explore the influence that they might have, we added an auxiliary pathway to the model (see below). We do not mean to suggest, in having added this pathway, that something as simple as our implementation underlies strategic processes; rather, we included it as a way of approximating the influence that we assume strategic processes would have on the time course of information processing.

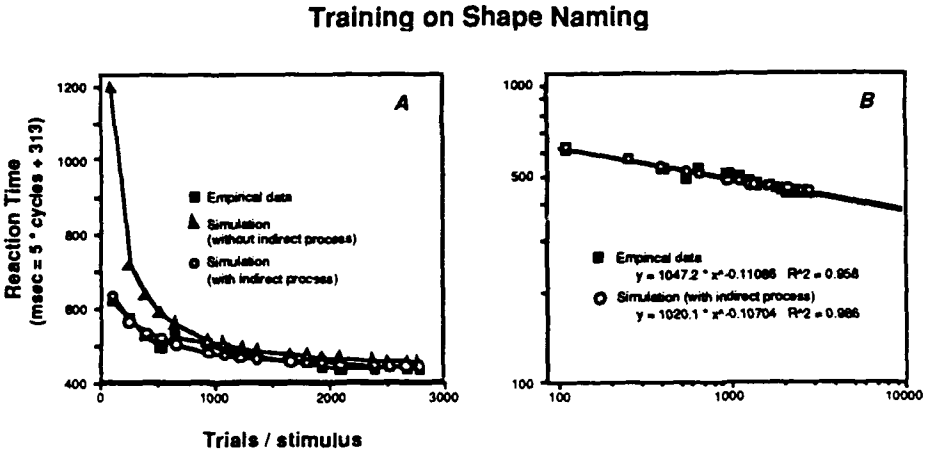


Figure 10. Training data for the shape naming task in Experiment 3 of MacLeod and Dunbar (1988), and the results of two simulations of this data. Panel A plots the empirical data, and the results of simulations with and without the indirect pathway for shape naming (see text). Panel B is a log-log plot of the empirical data and the results of the simulation using the indirect pathway for shape naming, with regression lines computed for each set of data independently.

The new pathway was comprised of connections from the intermediate units in the shape pathway to a new set of intermediate units in a separate module, and connections from this module to the model's output units (see Figure 11). We will call this new pathway the indirect pathway, to distinguish it from the usual "direct" pathways used by the network. The indirect pathway was meant to represent the involvement of a general purpose module (or even set of modules) that has been committed to the shape naming process for the current task. The connections in the indirect pathway were assigned a set of strengths that allowed it to be used for shape naming, before the effects of training had accrued in the direct pathway. This captured the assumption that such a mechanism can be rapidly programmed to perform a given task. Because the indirect pathway relied on an extra set of units, processing was slower than in the direct pathway. This conforms to the common assumption that processing relying on general purpose mechanisms is slower than automatic processing (e.g., Posner & Snyder, 1975).

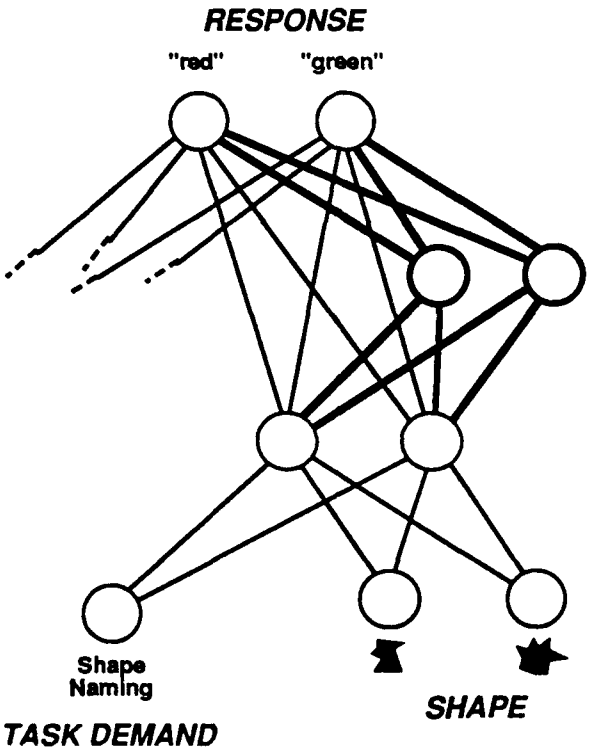


Figure 11. Detail of the pathways used for shape naming. Highlighted elements make up the indirect pathway.

The results of adding the indirect pathway to the network are shown in Figure 10a: the simulation's performance is now much closer to that of the subjects. Figure 10b shows that the best-fitting power functions for the empirical data and the simulation are almost identical. By comparing the model's performance with and without the indirect pathway, it can be seen that as training progresses, performance relies increasingly on the direct pathway. This is because, as the connection strengths in the direct pathway increase with training, processing in this pathway becomes faster. Our finding that such a transition from one processing mechanism to another follows a power law is similar to one described by Logan (1988), in which the transition from an algorithm-based to a memory-based process also produced a power law.

Interference effects. Figure 12 shows the interference and facilitation effects that were observed for the two tasks after 1, 5, and 20 days of practice on shape naming. Panel A shows the empirical data from MacLeod and Dunbar (1988; Experiment 4), and panel B the model's performance. Most importantly, we observe that the model captures the reversal of roles of shape naming and color naming. For both the subjects and the simulation, shape naming was initially much slower than color naming. Shape naming also showed interference and facilitation from colors early on, while color naming was not affected by shapes. At the final point in training, the relationship between the two processes reversed: shape naming became the faster process, while its sensitivity to interference was reduced and its ability to produce interference (with color naming) increased. At the intermediate point, the two processes were more comparable in their overall speed, and were able to influence each other.

In the model, shape naming started out as the slower process because, early in training, the strength of the connections in this pathway were still much smaller than those in the color pathway. The relationship between the two processes at this point was directly analogous to the relationship between word reading and color naming in Simulation 1. Note, however, that in this simulation color naming started out with the opposite role: initially, it was the process that was insensitive to interference or facilitation, and that was able to produce these effects. Color naming assumed this opposite role without any change in the strength of connections in its pathway. This makes it clear that the *absolute* strength of a pathway (i.e., the magnitude of connection strengths) is not the only relevant variable. *Relative* strength, compared with a competing pathway, is also important in determining whether or not a process will produce or be subject to interference in a Stroop-like task. This is further substantiated by the patterns of performance at the end of training. By this point, the strength of the shape pathway exceeded that of the color pathway. Accordingly, shape naming became faster than color naming, insensitive to colors, and able to facilitate and interfere with color naming.

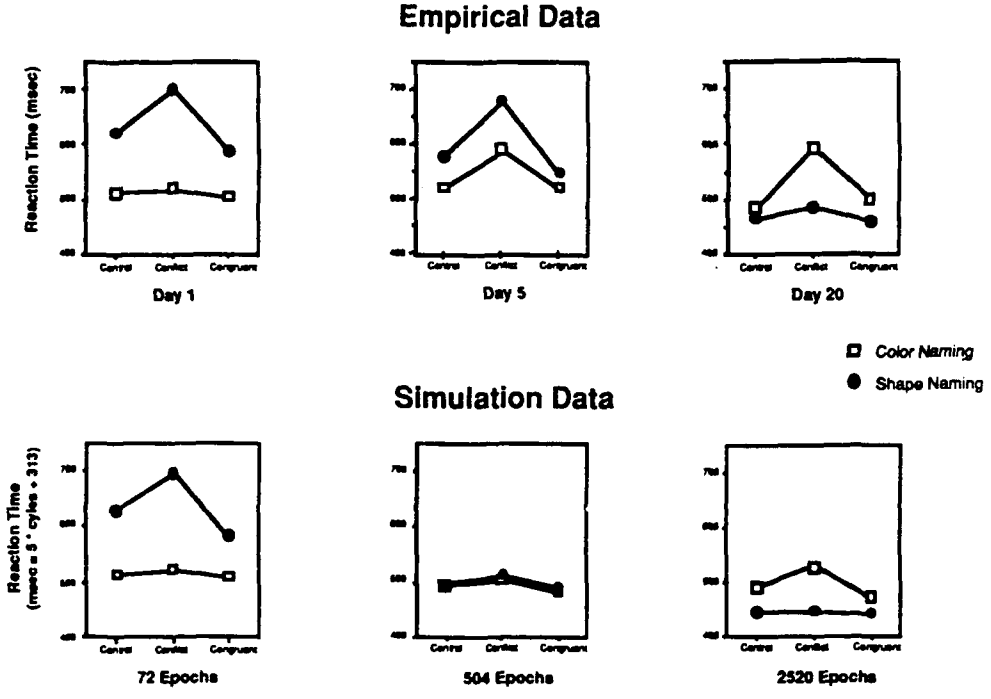


Figure 12. Interference and facilitation in the shape naming and color naming tasks after varying amounts of training on shape naming. Panel A shows the results obtained in MacLeod and Dunbar's Experiment 3 (1988), after subjects had received 1, 5, and 20 sessions of practice. Panel B shows the simulation results at the corresponding points in training.

Based on their data, MacLeod and Dunbar suggested that the Stroop effect can be understood in terms of the relative position of competing tasks along a continuum of automaticity, and that the position of the tasks along this continuum can be influenced by training. The results of this simulation are consistent with such a view, and demonstrate that the observed effects can be explained by increases in pathway strength that accompany training. This is an important finding, for it suggests that the same process can appear to be automatic (faster and able to influence a competing process) or controlled (slower and influenced by a competing process), depending upon the context in which it occurs.

There is, however, one respect in which the behavior of the model differs qualitatively from that of the subjects in MacLeod and Dunbar's study. This concerns the degree of interaction between processes that are of comparable strength. At the intermediate point in training, the empirical data show that each task interfered substantially with the other. In the simulation, though there was some mutual interference, the amount was rather small. This was a robust property of the model: processes of comparable strength showed less influence on one another than stronger processes did on weaker ones. Thus, it appears that additional factors outside the scope of our model may be involved when competing processes are of comparable strength. In fact, it is difficult to account for these findings even on other, more traditional grounds.[1] In this light, the mutual interference effect remains a general challenge to models of interference phenomena, and warrants further research.

Allocation of Attention

Simulation 5. Attention and Processing

A primary reason for studying interference effects it that they can tell us something about the requirements of different processes for attention. Thus, in the Stroop task it is assumed that information in the irrelevant channel is not attended to. To the extent that this unattended information can produce interference, it must not rely on attention to be processed. The lack of a requirement for attention is one of the primary criteria for automaticity (Posner & Snyder, 1975; Shiffrin & Schneider, 1977). It has often been assumed that automatic processes not only do not *require* attention, but also that they are

---

[1] For example, mutual interference might be thought to reflect an underlying probability mixture of trials involving unidirectional interference in each of the two directions. Thus, at an intermediate point in training, the shape pathway might interfere with the color pathway for some stimuli (or subjects), while the reverse is true for others. The effect of averaging over items (or subjects) would be that interference would *appear* to be bidirectional. However, the size of this average should be less than the amount of interference produced by colors early in training or by shapes late in training. This is because at the intermediate point only a subset of stimuli (or subjects) would be contributing to interference in each direction, whereas performance should be more homogeneous at the beginning and end of training. In fact, the data indicate that mutual interference was of roughly the same magnitude as the interference effects at the extremes of training. A probability mixture can not explain this finding.

not *influenced* by attention (e.g., Posner & Snyder, 1975). Kahneman and Treisman (1984) refer to this as the "strong automaticity" claim. They and others (e.g., Logan, 1980) have challenged this view, providing a large body of evidence that suggests that few processes, if any, occur entirely independent of attention (e.g., Kahneman & Chajczyk, 1983; Kahneman & Henik, 1981; Treisman, 1960). For example, Kahneman and Henik (1981) and Kahneman and Chajcyk (1983) showed that in the Stroop task the allocation of attention can influence the degree to which word reading interferes with color naming.

On the basis of these and related findings, Kahneman and Treisman (1984) have argued that automatic processes are subject to control by attention, although individual processes may differ in their degree of susceptibility to such control. Our model presents a view of automaticity that concurs with both of these points. In Simulation 1 we showed that while processing can occur in absence of attention — capturing the "involuntariness" of automatic processes — this autonomy was limited: even though words were processed without the allocation of attention, thereby interfering with color naming, they did not determine the response. Thus, even the strongest processes were controlled by attention. Furthermore, the model shows that control by attention is a matter of degree: this was seen in the gradual development of interference effects that occurred as strength of processing increased with training in Simulation 4.

In the following simulations, we examined the relationship between requirements for attention and strength of processing more directly. First, we looked at the effects that reducing attention had on performance of the color naming and word reading tasks. The amount of attention allocated to a task was represented as the activation value of the task demand unit associated with that task. Figure 13a shows reaction times to control stimuli in the word reading and color naming tasks, as a function of task demand unit activation for each of the corresponding processes. Two phenomena are apparent. For a given level of performance, color naming required more attention than word reading. However, *both* tasks were influenced by the allocation of attention. Even the word reading process showed degradation with reduced attention. Indeed, the fact that stronger pathways are controlled by attention is what allows the model to perform a task using the weaker of two competing pathways.

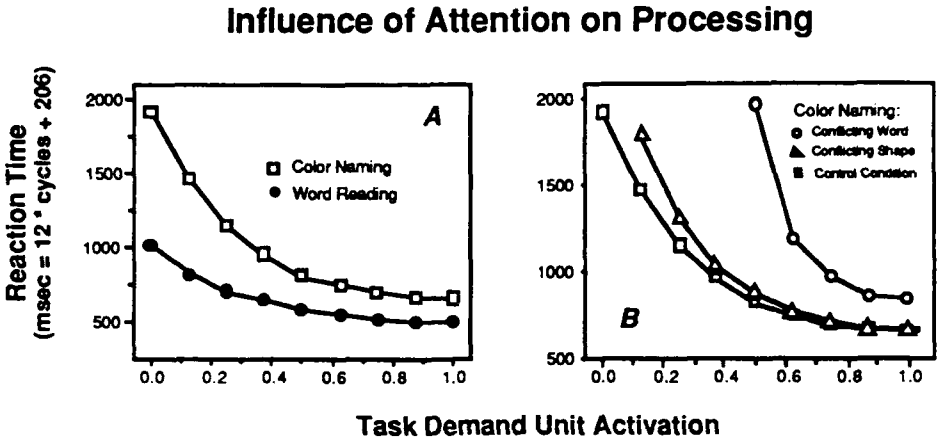


Figure 13. Influence of attention on processing. Panel A shows differences in the requirements for attention between color naming and word reading, and the effect on these two processes of reducing activation of the task demand unit. Panel B shows the different requirements for attention of the color naming process when it must compete with a stronger process (word reading) and a weaker one (shape naming, early in training).

Although stronger pathways rely less on attention, requirements for attention are influenced by more than just the absolute strength of a pathway; they are also affected by the circumstances under which a process occurs. Figure 13b graphs the requirements that color naming had for attention under three different conditions: no competing information, and conflicting information from a weaker process (shape naming, early in training) and a stronger process (word reading). In the two conflict conditions, color naming showed very different requirements for attention, depending upon the strength of the competing pathway. For a given level of performance, greater task demand unit activation was required for competition with the stronger process than with the weaker process.

The performance of the model under these conditions demonstrates that, although processing can occur in the absence of attention, all processes are affected by attention. Like the other attributes of automaticity, requirements for attention vary according to the strength of the underlying pathway, and the context in which the process occurs. The stronger a process is, the less are its requirements for attention, and the less susceptible it is to control by attention, increasing the likelihood that it will produce interference.

Simulation 6. Response Set Effects: Allocation of Attention at the Response Level

In the preceding simulations, we explored the role that attention plays in selecting information from one of two competing pathways. Attentional selection occurred at the level of the intermediate units, where information in the two pathways was still separate. However, the attention allocation mechanism used in this model is a general one, and can be applied to other levels of processing as well. In the following simulation, this mechanism is used to select a particular set of responses at the output level of the network. This provides an account for response set effects that have been observed in empirical studies (e.g. Dunbar, 1985; Klein, 1964; Proctor, 1978).

Response set effects reflect the fact that information related to a potential response leads to more interference (and facilitation) than information unrelated to the task. In the standard Stroop experiment, information in the irrelevant dimension is always related to a potential response. Potential responses are said to make up a "response set." For example, in the color naming task, when the word RED is written in green ink, although "red" is an incorrect response in that particular trial, it will be a correct response on other trials. Thus, both "red" and "green" are in the response set. However, if the color blue never appears, then the word BLUE is not in the response set, since it is never a response in the task. Several studies — using both the color naming task (e.g., Proctor, 1978) and a picture naming task (Dunbar, 1985) — have shown that words that are not potential responses produce significantly less interference than words that are in the response set.

An explanation that is commonly offered for this effect is that members of the response set are primed, either by instructions for the task (i.e, by informing the subjects of the stimuli they will have to respond to) or through experience with the stimuli in the course of the task itself (e.g., Kahneman & Treisman, 1984). The current model provides a related account of response set effects, in terms of the selective allocation of attention to members of the response set. The same mechanism that we used to allocate attention to a particular pathway in previous simulations can be used to allocate attention to a particular response or set of responses at the output level. In the previous simulations, allocation of attention to a processing pathway placed the intermediate units in that pathway on a more responsive part of their activation curve. This occurred through the activation of a task demand unit which offset the negative bias on intermediate units in that pathway. The same mechanism can be

implemented at the response level, by adding a negative bias to each of the output units, and having the allocation of attention to a response offset the negative bias on the appropriate output units.[1]

We simulated the response set effects observed for a picture naming task used in an experiment by Dunbar (1985). In this task, a word was placed in the center of a picture and subjects were required to name the picture and ignore the word. Subjects' performance in this task was almost identical to that in the standard color naming task: picture naming was slower than word reading, the word both interfered with and facilitated picture naming, and the picture had no effect on word reading (for similar studies, cf. Fraisse, 1969; Glaser & Dungelhoff, 1984; and Lupker & Katz, 1981) In Dunbar's experiment, there were five pictures of animals (horse, bear, rabbit, sheep, and cat[2]), and thus five possible responses. Some of the word stimuli used were potential responses (e.g., HORSE and BEAR), while others were animal words that were not in the response set (e.g., GOAT and DONKEY). Dunbar found that words in the response set produced significantly more interference than words which were not (see Table 2 below).

To simulate this experiment, the model used for Simulation 1 was extended in the following ways. First, we increased the number of units at each level of processing in each of the two pathways to ten (see Figure 14). One pathway was used to represent picture naming, while the other was used to represent word reading. Five output units were used to represent potential responses, and were labelled "horse", "bear", "rabbit", "sheep" and "cat". The remaining five were used to represent words that were not part of the response

---

[1] This mechanism was implicit at the output level in the previous simulations. To see this, imagine that a negative bias was associated with each of the output units, just as it was with the intermediate units. However, because both output units were in the response set, attention was maximally allocated to each. This would offset the negative bias on both of them. That is, the bias terms on the output units would always be equal to zero.

[2] To reflect the analogy between picture stimuli in this task and color stimuli in the classic Stroop task, we will refer to picture stimuli in lower case, as we have already done for color stimuli.

set , and were labelled "goat", "donkey", "dog", "mouse" and "seal." The ten input units in each pathway corresponded to these output units, and were labelled accordingly.[1]

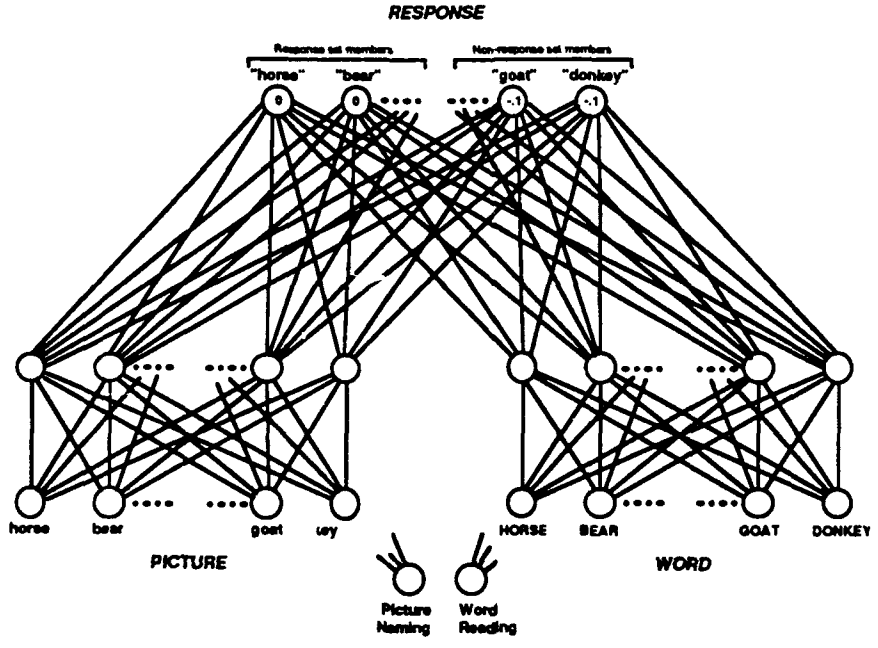


Figure 14. The network used to simulate response set effects. The numbers that appear inside each output unit are the bias terms that were assigned to these units during testing (see text).

The new network was trained in a way that was analogous to training in Simulation 1. Both pathways were trained on all ten of their inputs, with stimuli in the picture pathway receiving one tenth the amount of training received by those in the word pathway. During training, attention was allocated maximally to all of the units in both pathways, as was the

---

[1] The network was trained on input to the non-response set color units but, because the corresponding stimuli were not in the response set, they were never used in testing. These units were included strictly to maintain symmetry between the two pathways in the network, so that differences in processing between them could not be attributed to architectural asymmetries.

case in previous simulations. During testing, however, several adjustments were made in the allocation of attention. First, in order to simulate the somewhat smaller effects of words on picture naming (Dunbar, 1985) than on color naming (Dunbar & MacLeod, 1984), we increased the size of the attentional effect at the level of the intermediate units, much as we did in Simulation 2 (resting negative bias on all intermediate units of -4.5; strengths from the task demand units to intermediate units of 4.5).[1] In addition, attention was allocated differentially among the output units: non-response set units (e.g., "goat" and "donkey") were given a partial negative bias (-0.1). This corresponded to the hypothesis that, during testing, subjects allocate attention maximally to relevant responses and disattend to irrelevant responses, though not completely (cf. Deutsch, 1977; Kahneman & Treisman, 1984). The amount of negative bias applied to non-response items was chosen to capture the empirical data as accurately as possible. However, the fact that the size of the bias (-0.1) was smaller than the negative bias for intermediate units in the disattended pathway (-4.0) is consistent with empirical data demonstrating that selection by stimulus set is easier than selection by response set (cf. Broadbent, 1970; Kahneman & Treisman, 1984; Keren, 1976). The difference in bias between intermediate and output units is also consistent with the view that subjects are less able to allocate attention selectively to different representations within a module than to representations in different modules (e.g., Navon & Miller, 1987; Wickens, 1984).

---

[1] This difference does not seem to be due to a difference in strength between picture naming and color naming, since both have comparable reaction times in the control condition (approximately 650 msec). The comparability of naming times for colors and pictures was first noted by Cattell (1886).

Table 2. Response Set Effects.*

Condition	*Dunbar (1985)*	*Simulation*
Response set conflict	781	777
Non-response set conflict	748	750
Control	657	664
Congruent	634	628

* Empirical data are reaction times in milliseconds. Simulation data are cycles * 4.5 + 488.

Table 2 presents data from Dunbar's (1985) experiment as well as the results of the simulation. In both cases, stimuli that were not in the response set produced less interference than response set stimuli. In the model, this difference was due to the partial negative bias on the non-response set output units, which simulated failure to maximally attend to corresponding responses. This negative bias led to partial inhibition of these units, reducing the degree to which they were activated by input stimuli. As a result, they contributed less to competition within the response mechanism than output units that were in the response set. The simulation demonstrates that attention can be allocated to the response units using exactly the same mechanism that was used to allocate attention to a pathway. When attention is allocated using this mechanism standard response set effects are obtained.

General Discussion

We have shown that the mechanisms in a simple network-based model can explain many of the phenomena associated with attention and automaticity. With regard to the Stroop effect, the model shows that these mechanisms can capture a wide variety of empirical effects. Among these are the asymmetry of interference effects between word reading and color naming; the fact that interference effects are typically larger than facilitation effects (Dunbar & MacLeod, 1984); that presenting the color before the word produces substantially less interference than would be expected simply from differences in the speed of processing

(Glaser & Glaser, 1982); and that words which are not in the response set produce less interference with color naming than words which are (Dunbar, 1985; Klein, 1964). In addition, the model exhibited many of the phenomena associated with the development of automaticity, including reductions in reaction time and variance that follow a power law (Logan, 1988; Newell & Rosenbloom, 1980); gradual development of the ability to produce interference accompanied by a reduction in susceptibility to interference (MacLeod & Dunbar, 1988); and a reduction of the requirements for attention as learning occurs (Logan, 1978; Shiffrin & Schneider, 1977).

The model provides a common explanation for these findings in terms of the strength of processing pathways. This account goes beyond many other theories of automaticity by describing an explicit set of processing mechanisms from which the empirical phenomena are shown to arise. These mechanisms provide a basis for learning, the time course of processing, and the influence of attention. Several important features of automaticity emerge from this account, including the fact that the properties of automaticity are continuous, and that their emergence depends largely on the strength of a process *relative to* the strengths of competing processes.

The model is not perfect in its present form. For example, it does not account for the fact that presenting a word sufficiently in advance of the color reduces interference (Glaser & Glaser, 1982). It also shows less interaction between processes of comparable strength than the available data seem to indicate (MacLeod & Dunbar, 1988). Some of these shortcomings may be due to the fact that the model does not include mechanisms for the processing of strategic components in a task (e.g., interpretation of task demands, evaluation of the response set, compensation for a preceding conflicting stimulus, etc.). Further research and development are needed in order to capture these and other aspects of performance. Nevertheless, the successes of the model to date indicate the usefulness of the general approach. In the remainder of this discussion we consider the implications of the approach for issues beyond those directly addressed in our simulations.

Reconsidering Controlled and Automatic Processing

The model demonstrates that differences in interference effects are not sufficient to make a distinction between different *types* of processes. Often it has been assumed that if one process interferes with another, the process that produces interference is automatic and the

other is controlled. However, the model shows that this disparity can be explained by differences in the *strength* of two processes that use qualitatively identical mechanisms. Furthermore, both the model and recent empirical evidence demonstrate that the same process can — according to interference criteria — appear to be controlled in one context and automatic in another.

In this respect, our model provides a very different account of the Stroop effect than other models, such as one described by Hunt and Lansman (1986; also see Reed & Hunt, 1986). Their model is based on a production system architecture that is a modified version of the one used in ACT* (Anderson, 1983). Their model distinguishes between controlled and automatic processing based on the manner in which one production influences the firing of others. In controlled processing, one production activates a representation in working memory that matches the criteria for another, increasing the activation value for that production, and hence its likelihood of firing. In automatic processing, however, productions can influence each other without relying on working memory, through the direct spread of activation from one production to another. In the Hunt and Lansman model, color naming is assumed to be a controlled process, which relies on working memory. As such, it is highly influenced by the contents of working memory. In particular, in the conflict condition, when there is competing word information in working memory, processing is slowed resulting in Stroop-like interference. In contrast, word reading is assumed to be automatic, and thus to occur largely through the direct spread of activation between production rules. This makes it less susceptible to influence by the contents of working memory.

The Hunt and Lansman model provides an explicit account of the nature of controlled processing (in terms of working memory and productions), which our model does not. However, it faces serious limitations. First, it fails to capture some of the basic features of the Stroop task: in the control conditions, color naming and word reading are the same speed. Furthermore, the color can interfere with and facilitate word reading, neither of which occur in empirical studies. It is not clear whether these failures to fit the empirical data result from fundamental limitations in their overall approach or the particular implementation they report. Most importantly, however, this model accounts for differences in interference effects between color naming and word reading by assuming that these represent different types of processes: one is controlled while the other is automatic. By making such qualitative distinctions, it seems that this approach can not — in principle

— account for the MacLeod and Dunbar findings, which show that color naming can appear to be controlled in one context but automatic in another (i.e., when it is in competition with a less practiced task). In contrast, our model accounts for these findings in terms of differences in the relative strengths of two competing pathways, both of which might be considered to be automatic.

Although our model indicates that there is a continuum of automaticity, it does not reject the existence of controlled processing. At the extreme low end of the automaticity continuum, where there is no pre-existing pathway to perform a task, it is clear that processing must occur in a very different way. Consider, for example, a subject who is told to say "red" when a particular random figure is presented, and to say "green", "blue", etc. for each of several other shapes. Initially, the subject will lack the relevant connections for performing the task. At this point, the task might be performed with the assistance of the experimenter (e.g., by being reminded of the color word that corresponds to the shape on the screen). The subject might try to learn each correspondence, using verbal associations to the shapes (e.g., orange is the name of the shape that looks like Florida) or other mnemonics. We assume that such processes rely on indirect pathways that can be used to establish at least a few arbitrary associations relatively quickly; but that processing in such pathways is slow and requires effort to maintain. At the same time, as practice progresses and the subject receives feedback regarding responses, connections would be starting to build in a pathway that will ultimately allow the shape to directly activate the correct response, without recourse to indirect verbal and/or mnemonic mediation. While learning occurs more gradually in such direct pathways, it leads to processing that is faster and stronger than is possible using indirect pathways.

Thus, while subjects can respond correctly without external help after only a few trials, we assume that the direct pathway would generally not at this point be sufficient to produce the response. The activation of a response would be based on the combination of information from both the direct and indirect mechanisms, with the relative importance of the direct pathway growing steadily over trials, and the contribution made by the indirect pathway diminishing. In summary, what we see during the early phases of practice may reflect a gradual transition from a reliance on indirect to direct pathways.

There is a partial correspondence between our direct/indirect distinction and the traditional distinction between controlled and automatic. As already noted, a process based on indirect pathways would have all the earmarks of what is typically called a controlled process: it

would be slow, it might consist of a series of steps which can be disrupted or interfered with, and it might depend on declarative (verbal) memory (e.g., "Florida is orange"), or other explicit mnemonics requiring effort and the allocation of attention. On the other hand, at high extremes of practice, direct performance would correspond closely to what typically has been called automatic: processing would be much faster, less susceptible to interference, more capable of producing interference, and less influenced by the allocation of attention. In between, however, the correspondence between these distinctions breaks down. As the simulations presented in this paper demonstrate, a process that is completely direct can — under some circumstances — exhibit all of the properties usually ascribed to a controlled process. Thus, we propose that processes which have previously been classified as controlled might more profitably be segregated into those which are direct and those which are indirect. Within the range of direct processes, there would be a continuous spectrum of pathway strengths that span the range of different degrees of automaticity. Table 3 illustrates the correspondence between traditional usage and the terms proposed here.

Table 3. Types of Processing

Novel Tasks			Highly Practiced Tasks
Indirect →	(Indirect with very weak direct) →	Weak Direct →	Strong direct
Controlled →		Controlled →	Automatic

The model provides an explicit account of direct processes, and shows how changes in the strength of these processes — which result from practice — can lead to seemingly qualitative changes in performance. The model is less explicit about indirect processes. Because our focus was on the nature and interaction of direct processes, indirect processes were included in only one simulation (Simulation 4), in order to capture performance of the shape naming task early in training.

The significant aspects of our implementation of an indirect process (see Simulation 4) were that it relied on an extra module in the processing pathway, and that the connections in this pathway were available early in training, before connections had developed in the direct pathway. This implementation captured the slower dynamics of indirect processes that commonly have been observed . However, it did not capture other features of indirect processes, such as their flexibility, and the general purpose nature of the mechanisms involved. Nevertheless, there are extensions of our model that might capture some of these features. For example, it would be possible to replace the single additional module used in Simulation 4 with a series of modules — perhaps participating in a number of different processing pathways — that had highly adaptive but quickly decaying connections. These properties would capture the flexible, general purpose, but slower and less stable nature of indirect processing. While PDP research in this area is just beginning, a number of PDP models that use such mechanisms have already begun to appear (e.g., Schneider, 1985; Schneider & Detweiler, 1987; Hinton & Plaut, 1987).

Attention and the Control of Processing

We have argued that an important difference between our direct/indirect distinction and the traditional dichotomy between controlled and automatic processing is that in our distinction, processes of either type can exhibit performance characteristics traditionally associated with controlled processing — such as slower speed and susceptibility to interference. The same difference can be found between these two approaches concerning their claims about the attentional control of processing.

At the heart of the theoretical distinction between controlled and automatic processing are two basic assumptions: controlled processing depends on the allocation of attention; automatic processing occurs independently of attention. As we discussed in Simulation 5, there is reason to believe that few, if any processes are entirely immune to the affects of attention. In the simulations presented here, even the strongest pathways — in which processing exhibited all of the other attributes of automaticity — processing was affected by the allocation of attention. For example, in Simulation 1, although processing in the word pathway occurred without the allocation of attention — leading to interference with color naming — this processing was only partial, and was insufficient to determine which

response was made. Simulation 5 showed that the word reading process was directly influenced by changes in the allocation of attention.

The proposal that direct processes are subject to attentional control is quite different from the proposal that a particular *task* is subject to attentional control. Thus, others (e.g., Shiffrin, in press) have attempted to explain the fact that automatic processing tasks such as word reading are subject to attentional control by arguing that behavior relies on numerous processes, some of which are automatic, and some of which may be controlled. In this view, control over the performance of a task could be explained by the allocation or withdrawal of attention from the controlled processes involved, preserving the independence of the automatic processes from the effects of attention. While we do not dispute the claim that behavior is composed of many component processes, our model asserts that all of these processes may be subject — in varying degrees — to control by attention.

Attention as the Modulation of Processing

Given that all cognitive processes are subject, in some degree, to attentional control, the question arises as to how this control is achieved. Attention is implemented in the model as the modulation of processing in a pathway. This occurs by input from attention (task demand) units, which cause a shift in the responsiveness of units in a processing pathway.

Attention uses exactly the same processing mechanisms as the other components of the model. The connections from the attention units to the units in a processing pathway are of the same type as the connections within the pathway itself, and attentional information is represented in the same way as any other information in the network: as a pattern of activation over a set of units. As such, the input that a pathway receives from the attention units is qualitatively the same as input received from any other source of information in the network. Attention can be viewed simply as an additional source of information which provides a sustained context for the processing of signals within a particular pathway. Thus, attentional mechanisms are not given special status in the model and, in general, an attentional module can be thought of as any module that has a set of connections that allow it to modulate processing in another pathway. There may be many such modules within a system, a given module may modulate one or many pathways, and it might even participate directly in one set of processing pathways, while it serves to modulate others. This view

of attentional control is similar to the "multiple resources" view that has been expressed by others (e.g., Allport, 1982; Hirst & Kalmar, 1987; Navon & Gopher, 1979; Wickens, 1984).

Our implementation of attentional modulation is different from several other recent accounts of attention. In Anderson's ACT* theory (1983), attention is related to the competition for representation in working memory rather than the modulation of processing in otherwise automatic pathways. Schneider (1985) provides an account of attention as the modulation of information in a PDP network. However, his model uses a mechanism for attentional modulation (multiplicative connections) that is qualitatively distinct from other types of processing in the network, unlike the model we have presented.

The notion of attention as a modulator, together with the idea that processing is continuous and that the resulting activations are graded in strength, has a long history in the attention literature; the idea is essentially the same as that suggested by Treisman (1960). Treisman claimed that messages outside of the focus of attention were not completely shut out; rather, the flow of information was simply "attenuated" on the unattended channel. This is exactly what happens in our model. Indeed, the very same mechanisms of pathway modulation that we have used to implement task selection in the Stroop task (color or word) could be used to implement channel selection in dichotic listening (e.g., Treisman, 1960), spatial allocation of attention (e.g., Kahneman & Henik, 1981), category search (e.g., Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977), or other tasks involving selective attention.

A number of these phenomena have begun to be explored productively using the PDP framework (e.g., Mozer, 1988; Phaff, 1986; Schneider & Detweiler, 1987). For example, Mozer (1988) has addressed the spatial allocation of attention using PDP mechanisms similar to the ones we have described. By introducing a network of units corresponding to retinal locations (rather than feature dimensions), and having attention bias units representing particular locations, attention can be allocated to specific locations. As in our model, information that is in the focus of attention is processed to a greater degree than information that is not in the focus of attention. Extension of our model along these lines would provide a means for simulating spatial allocation of attention in the Stroop task (e.g., Kahneman & Henik, 1981).

Finally, with regard to the mechanisms of attentional modulation, an important general issue is whether attention facilitates processing within the attended pathway, suppresses it

in the unattended pathway, or both. Our framework is, in principle, agnostic on this issue; attention could be implemented either as a facilitative or an inhibitory effect, or a combination of both. In the simulations we presented, however, attention had primarily a facilitative effect: task specifications put units in the attended pathway in a more responsive portion of their dynamic range. This succeeded in capturing the central phenomena of the Stroop effect, and the relationship between practice, automaticity and attention that we set out to explain. However, there are reasons to suspect that effort is also required to filter out potentially interfering messages. For example, processing is typically slower on control trials if these are mixed with interfering trials. One way to account for this is to assume that attention is required both to suppress the unattended channel and to enhance processing in the attended one, and that suppressing one requires resources that take away from the ability to facilitate the other. The modelling framework we have described could be used to explore — in simulations — which of these explanations can best account for the relevant empirical data.

Continuous Nature of Processing

The assumption that information is graded and is propagated continuously from one level to the next distinguishes our model from discrete stage models, in which processing must be complete at one stage before information becomes available to others. In the model presented here, information at one level is continuously available to subsequent levels. As such, a process need not be completed in order for it to affect performance. It is precisely the partial processing of information in the stronger of two pathways that produces interference and facilitation effects.

In this respect our model is similar to one proposed by Logan (1980). Both models make use of continuous processing mechanisms, and explain interference (and facilitation) effects in terms of the relative strength of the pathways used by competing processes. According to Logan's model, "evidence is assumed to accumulate over time in some composite decision process until a threshold is exceeded and a response is emitted" (p. 528, Logan, 1980). Different sources of evidence (e.g., different stimuli, or different stimulus dimensions) are weighted so that evidence accumulates from each at different rates. This model accounts for the Stroop effect by assigning stronger weights to word reading than to color naming. The strength of the connections in a pathway in our model are analogous to the weights assigned to a process in Logan's model. In both models, attention acts by

modulating the effectiveness with which information accumulates from each process in a manner that is responsive to the demands of the current situation.

However, our model differs from Logan's in several important respects, some of which lead to significant differences in performance. First, Logan's model is a linear model, with respect to the way in which information accumulates both as a function of time and as a function of pathway strength. The processing mechanisms in our model are non-linear in both of these respects. This allows it to account for the asymmetry between interference and facilitation. Logan's model can not account for this finding. However, perhaps the most important difference between the two models is that Logan's model does not include any mechanisms for learning. The weights associated with automatic processes are fixed. One of the primary strengths of our model is that it can directly address the relationship between training and automaticity in terms of an integrated set of learning and processing mechanisms.

Strength and Instance-based Accounts of Automaticity

The model presented in this paper is based on the assumption that direct processes develop through the strengthening of connections between processing units. In this respect, it is one of a general class of models that explain learning in terms of a strengthening process. An alternative approach to learning and automaticity is instance based (e.g., Hintzman, 1986; Logan, 1988). According to instance theory, each exposure to a stimulus is encoded separately in memory. In Logan's model, both encoding as well as the retrieval of stimulus-related instances is obligatory. Retrieval times for individual instances are normally distributed, with the first instance retrieved controlling the response. Logan has demonstrated that instance theory accurately predicts practice effects in non-conflict tasks, accounting for the fact that both the mean and the standard deviation of reaction times decrease according to a power law with the number of trials (i.e., instances encoded). The theory also predicts that the exponent for both functions should be the same.

A number of strength-based accounts have provided fits to the power law for mean reaction time (e.g., Anderson, 1983; Schneider, 1985), and the model we presented satisfies the additional constraint that standard deviation decrease at the same rate as mean reaction time. However, Logan (1988) has expressed other concerns about strength-based theories of learning. For example, he claims that strength-based accounts must rely on fixed

prototypes: it is the strengthening of the connection between "generic stimuli" and "generic responses" that constitute learning in such systems. This criticism applies primarily to theories using discrete, or "local" representations of stimuli. Elsewhere (McClelland & Rumelhart, 1985) it has been shown that this limitation of the strength-based approach can be overcome with the use of distributed representations. In such systems, memory for an event is not encoded in a single connection between a generic stimulus and a generic response, but in the strengths of a *set* of connections involving a number of different units which are used to provide overlapping but nevertheless distinct representations of individual stimuli and responses. In the current model, local representations were used. However, in preliminary investigations using distributed representations we have had no difficulty reproducing the basic interference phenomena (resulting from unequal amounts of training) reported in this paper. These effects appear to be general to cascaded PDP networks composed of continuous non-linear processing units.

For the moment, it appears that both instance- and strength-based theories are equally able to explain learning behaviors. However, it is not clear how an instance-based account will explain some of the interference phenomena we have addressed. As an assumption of his theory, Logan states that "attending to a stimulus is sufficient to retrieve from memory whatever has been associated with it in the past" (Logan, 1988, p.493). That is, retrieval is obligatory and, by implication, unmodulated. Interference is produced in this system when retrieved information associated with a stimulus conflicts with the desired response, and this information is retrieved before the relevant information. As with the simple speed of processing account, this suggests that given sufficient time for retrieval, colors should interfere with word reading as much as words do with color naming. However, we know from Glaser and Glaser's (1982) data that this is incorrect. Thus, instance theory faces the same difficulties that simple speed of processing accounts face in explaining Stroop interference effects.

More generally, as instance theory is currently developed, it does not specify mechanisms for attentional influences on behavior. Logan (1988) suggests that "the retrieval process can be controlled by manipulating retrieval cues or stimulus input, or both, and the subsequent decision process can be inhibited before it results in an overt response" (Logan, 1988, p. 513). However, no mechanism is provided for these processes, nor has it been demonstrated whether these processes can account for the interactions between interference and practice effects that we have addressed in this paper. While we have not provided a

mechanism which determines how and where attention will be allocated, we have specified a mechanism by which the allocation of attention can influence processing, and we have shown how this mechanism interacts with learning.

Resources and Capacity

A final issue concerns the notions of processing resources and capacity limitations. The model instantiates processes similar to the "multiple resources" view that has been expressed by others (e.g., Allport, 1982; Logan, 1985; Navon & Gopher, 1979; Wickens, 1984), and to the notion of functional cerebral distance described by Kinsbourne and Hicks (1978). These theories share the view that performance of a task typically involves a number of different processes, which in turn depend on a multiplicity resources. They predict that two behaviors will compete for processing capacity, and may interfere with one another to the extent that they rely on the same resources for different purposes. Our approach suggests ways in which we can begin to think, in more specific terms, about the nature of these resources, and how limitations in their capacity can affect performance. Thus, the modules that make up processing pathways can be thought of as a set of resources within the system. These resources are shared by two or more processes to the extent that their pathways intersect — that is, they rely on a common set of modules. In the model, when two signals to be processed by a particular module are disparate (i.e, they involve different patterns of activation), they will compete for representation within that module. In this sense, the processing capacity of that module, or resource can be thought of as being limited — that is, it cannot fully support the processing of both signals at once. Schneider (1985; Schneider & Detweiler, 1987) has presented a similar view of capacity limitations in terms of cross-talk within modules (also see Navon & Miller, 1987).

Although we have not pursued a quantitative analysis of capacity in this paper (see Rosenfeld & Touretsky, 1987 for an example of how this can be done in PDP systems), the model showed that when information from two sources converged on a common module (the response module of the network), interference occurred. We have obtained similar results in other simulations, in which two stimuli (e.g, two words) processed concurrently within the same pathway also led to interference.

As in the multiple resources view, our account of interference effects focuses on the capacity limitations of modules directly involved in a processing pathway; that is, modules

which lie in the pathway along which information flows from input to output. However, the model also suggests ways in which to think about other types of capacity limitations. For example, the attention module did not lie directly along one of the processing pathways in the network. Nevertheless, it played an important role in processing: Simulation 5 showed that all processes relied, to a greater or lesser extent, on the allocation of attention. For a given process, this required that a particular pattern of activation be present in the attention module. This pattern was different for different processes, so that any attempt to specify more than one process would lead to competition of representations within the attention module. From this perspective, the capacity of the attention module can be seen as limited — it may not always be possible to allocate attention maximally to all processes at once. Because stronger processes have weaker requirements for attention (see Simulation 5), such processes may be less susceptible to capacity limitations in the corresponding attentional module. This is consistent with the traditional notion that automaticity is associated with greater independence from capacity limitations of attentional resources. However, our approach allows that there may be more than one attentional resource (module) within the system, and that different processes may rely on different such modules. As such, the extent to which limitations in attentional capacity will affect performance will depend on the particular processes involved in performance of the task (or set of tasks), the extent to which these processes rely on attentional resources, and whether the attentional resources are the same or different for the various processes involved.

The significance of different sources of capacity limitations (e.g., those arising directly within a pathway, or within associated attentional modules) are in need of further clarification, both theoretically and empirically. However, we would like to re-emphasize, in this context, that attentional information is not qualitatively different from other information in our framework. The competition between patterns of activation within an attentional module is analogous to the competition that can occur between patterns of activation in any other module. This suggests that modulatory resources, such as the attentional module in our model, may be governed by the same sorts of principles and constraints that govern more local resources within the system.

Conclusion

The model that we have presented provides not only an account of the empirical data on the Stroop effect, but also a more general model of processing in highly practiced tasks and its

relation to attention. Like other theorists (e.g., Kahneman & Treisman, 1984; Logan, 1980; Schneider, 1985), we have noted that there are many problems with an all or none view of automaticity. Our model suggests that a more useful approach is to consider automaticity in terms of a continuum based on strength of processing. We have outlined a set of mechanisms that can produce gradual and continuous strengthening, and we have shown how these mechanisms can account for a variety of empirical phenomena concerning automaticity. In particular, these mechanisms capture the continuum that appears to exist in the attributes of automaticity, and relate this continuum directly to the effects of practice. Differences in practice lead to differences in the strength of processing, and this makes it possible to capture asymmetries of performance such as those observed in the Stroop task. The model also makes the point that Stroop-like effects can arise from the competition between two qualitatively similar processes — which differ only in their strength — questioning the traditional view that interference effects can be used reliably to distinguish between controlled and automatic processes. Finally, the model suggests ways in which it may be possible to characterize the notion of capacity in greater detail than has occurred up to now.

The mechanisms used in this model show how the principles of continuous processing — expressed in terms of the PDP framework — can be applied to the study of attention and the control of what we have called direct processes. A challenge for our model, however, and for the PDP approach in general is to characterize the mechanisms underlying indirect processing. We see this as an important direction for future research.

Authors' Notes

A preliminary version of this work was presented at the Annual meeting of the Psychonomics Society in Seattle, November 1987.

We gratefully acknowledge the helpful comments and suggestions made by David Galin, Gordon Logan, Colin MacLeod, David Rumelhart, Walter Schneider and David Servan-Schreiber.

This work was supported by a research grant from the Scottish Rite Schizophrenia Research Program, N.M.J., U.S.A. and a NIMH Physician Scientist Award (MH00673) to the first author; by funding from the Natural Sciences and Engineering Council Canada (OGP0037356) and the Department of Psychology of Carnegie Mellon University to the second author; and by ONR Contract N00014-82-C-0374, NR442a-483, and a NIMH Research Scientist Career Development Award (MH00385) to the third author.

Correspondence concerning this paper should be addressed to Jonathan D. Cohen, Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213.

References

Allport, D. A. (1982). Attention and performance. In G. I. Claxton (Ed.), *New directions in cognitive psychology*. London: Routledge and Keagan-Paul.

Allport, D. A., Antonis, B. & Reynolds, P. (1972). On the division of attention: A disproof of the single-channel hypothesis. *Quarterly Journal of Experimental Psychology, 24*, 225-235.

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review, 89*, 369-406.

Anderson, J. R. (1983). *The Architecture of Cognition*. Cambridge, MA: Harvard University Press.

Blackburn, J. M. (1936). Acquisition of skills: An analysis of learning curves. IHRB Report No. 73.

Broadbent, D. E. (1970). Stimulus set and response set: Two kinds of selective attention. In D.I. Mostofsky (Ed.), *Attention: Contemporary theories and Analysis*. New York: Appelton-Century-Crofts.

Brown, W. (1915). Practice in associating names with colors. *Psychological Review, 22*, 45-55.

Bryan, W. L. & Harter, N. (1899). Studies of the telegraphic language. The acquisition of a hierarchy of habits. *Psychological Review, 6*, 345-375.

Cattell, J. M. (1886). The time it takes to see and name objects. *Mind, 11*, 63-65.

Deutsch, J. A. (1977). On the category effect in visual search. *Perception & Psychophysics, 21*, 590-592.

Dunbar, K. (1985). The roles of multiple sources of interference in a picture-word analogue of the Stroop task. Unpublished PhD thesis, University of Toronto.

Dunbar, K., & MacLeod, C. M. (1984). A horse race of a different color: Stroop interference patterns with transformed words. *Journal of Experimental Psychology: Human Perception and Performance, 10*, 622-639.

Dyer, F. N. (1973). The Stroop phenomenon and its use in the study of perceptual, cognitive, and response processes. *Memory and Cognition, 1*, 106-120.

Fraisse, P. (1969). Why is naming longer than reading? *Acta Psychologica 30*:96-103.

Gatti, S. V. & Egeth, H. E. (1978). Failure of spatial selectivity in vision. *Bulletin of the Psychonomic Society, 11*, 181-184.

Glaser, W. R. & Dungelhoff, F-J. (1984). The time course of picture-word interference. *Journal of Experimental Psychology: Human Perception and Performance 10:*640-654.

Glaser, M. O., & Glaser, W. R. (1982). Time course analysis of the Stroop phenomenon. *Journal of Experimental Psychology: Human Perception and Performance, 8, 6,* 875-894.

Gumenik, W. E. & Glass, R. (1970). Effects of reducing the readability of the words in the Stroop color-word test. *Psychonomic Science, 20,* 247-248.

Hasher, L., & Zacks, R. T. (1979). Automatic and effortful processes in memory. *Journal of Experimental Psychology: General, 106,* 356-388.

Hinton, G. E. & Plaut D. C. (1987). Using fast weights to deblur old memories. In *Program of the Ninth Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Lawrence Erlbaum.

Hintzman, D. L. (1986). "Schema Abstraction" in a multiple-trace model. *Psychological Review 93:*411-428.

Hintzman, D. L., Carre, A., Eskridge, V. L., Owens, A. M., Shaff, S. S., & Sparks, M. E. (1972). "Stroop" effect: Input or output phenomenon? *Journal of Experimental Psychology, 95,* 458-459.

Hirst, W. & Kalmar, D. (1987). Characterizing attentional resources. *Journal of Experimental Psychology: General, 116, 1,* 68-81.

Hunt, E., & Lansman, M. (1986). Unified model of attention and problem solving. *Psychological Review, 93,* 446-461.

James, W. (1890). Principles of Psychology. New York: Henry Holt and Co.

Kahneman, D., & Chajczyk, D. (1983). Tests of the automaticity of reading: Dilution of Stroop effects by color-irrelevant stimuli. *Journal of Experimental Psychology: Human Perception and Performance, 9,* 497-509.

Kahneman, D. & Henik, A. (1981). Perceptual organization and attention. In M. Kubovy & J.R. Pomerantz (Eds.), *Perceptual organization.* Hillsdale, N.J.: Erlbaum.

Kahneman, D., & Treisman, A. (1984). Changing views of attention and automaticity. In R. Parasuraman, R. Davies, & J. Beatty (Eds.), *Varieties of attention.* New York: Academic Press.

Keren, G. (1976) Some considerations of two kinds of selective attention. *Journal of Experimental Psychology: General, 105:* 349-374.

Kinsbourne, M. & Hicks, R. E. (1978). Functional cerebral space: A model for overflow, transfer and interference effects in human performance. In J. Requin (Ed.), *Attention and Performance VII.* Hillsdale, NJ: Lawrence Erlbaum.

Klein, G. S. (1964). Semantic power measured through the interference of words with color naming. *American Journal of Psychology, 77,* 576-588.

Kolers, P. A. (1976). Reading a year later. *Journal of Experimental Psychology: Human Learning and Memory, 2,* 554-565.

LaBerge, D. & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology, 6,* 293-323.

Link, S. W. (1975). The relative judgement theory of two choice response time. Journal of Mathematical Psychology, 12, 114-135.

Logan, G. D. (1978). Attention in character classification: Evidence for the automaticity of component stages. *Journal of Experimental Psychology: General, 107,* 32-63.

Logan, G. D. (1979). On the use of a concurrent memory load to measure attention and automaticity. *Journal of Experimental Psychology: Human Perception and Performance, 5,* 189-207.

Logan, G. D. (1980). Attention and automaticity in Stroop and priming tasks: Theory and data. *Cognitive Psychology, 12,* 523-553.

Logan, D. G. (1985). Skill and automaticity: Relations, implications, and future directions. *Canadian Journal of Psychology, 39, 2,* 367-386.

Logan, D. G. (1988). Toward an instance theory of automatization. *Psychological Review, 95, 4,* 492-527.

Lupker, S. J. & Katz, A. N. (1981). Input, decision, and response factors in picture-word interference. *Journal of Experimental Psychology: Human Learning and Memory 7:*269-282.

MacLeod, C. M. (1989). Half a century of research on the Stroop effect: A critical review. Manuscript submitted for publication.

MacLeod, C. M. & Dunbar K. (1988). Training and Stroop-like interference: Evidence for a continuum of automaticity. *Journal of Experimental Psychology, 14,* 126-135.

McClelland, J. L. (1979). On the time-relations of mental processes: An examination of systems of processes in cascade. *Psychological Review,* 86,287-330.

McClelland, J. L. (in press). Parallel distributed processing: Implications for cognition and development. In Morris, R.G.M. (ed), *Parallel Distributed Processing: Implications for Psychology and Neurobiology.* Oxford: Oxford University Press.

McClelland, J. L. & Rumelhart, D.E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General, 114, 2,* 159-188.

Moray, N. (1960). Attention: Selective processes in vision and hearing. London: Hutchinson.

Mozer, M. (1988). A connectionist model of selective attention in visual perception. In the proceedings of the tenth annual conference of the Cognitive Science Society.Hillsdale, NJ: Erlbaum, pp. 195-201.

Navon, D. & Gopher, D. (1979). On the economy of the human processing system. *Psychological Review, 86,* 214-255.

Navon, D. & Miller, J. (1987) Role of outcome conflict in dual-task interference. *Journal of Experimental Psychology: Human Perception and Performance, 13:* 435-438.

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General, 106,* 226-254.

Neumann, O. (1980). *Informationsselektion und Handlungssteuerung.* Unpublished doctoral dissertation, University of Bochum, FRG. Cited in Phaff (1986).

Neumann, O. (1984). Automatic processing: A review of recent findings and a plea for an old theory. In W. Prinz and A.F. Sanders (Eds.), *Cognition and motor processes.* Berlin: Springer-Verlag.

Phaff, R. H. (1986). *A connectionist model for attention: Restricting parallel processing through modularity.* Unpublished doctoral dissertation, Unit of Experimental Psychology, University of Leiden, Netherlands.

Pillsbury, W. B. (1908). Attention. New York: Macmillan.

Posner, M. I. (1975). Psychobiology of attention. In M. S. Gazzaniga & C. Blakemore (Eds,), Handbook of Psychobiology. New York: Academic Press.

Posner, M. I., & Snyder, C.R. (1975). Attention and cognitive control. In R.L. Solso (Ed.), *Information processing and cognition.* Hillsdale, NJ: Erlbaum.

Proctor, R. W. (1978). Sources of Color-word interference in the Stroop color-naming task. *Perception and Psychophysics, 23,* 413-419.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review,* 85, 59-108.

Regan, J. (1978). Involuntary automatic processing in color naming tasks. *Perception and Psychophysics, 24,* 130-136.

Reed, P. & Hunt, E. (1986). A production system model of response selection in the Stroop paradigm. Unpublished manuscript.

Rosenfeld R. & Touretsky, D. S. (1987). Scaling properties of coarse-coded symbol memories. In *Proceedings of IEEE Conference on Neural Information Processing Systems-Natural and Synthetic.*

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986). Learning internal representations by error propagation. In D.E. Rumelhart, J.L. McClelland, and the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 1.* Cambridge, MA: MIT Press.

Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. (1986). A general framework for parallel distributed processing. In D.E. Rumelhart, J.L. McClelland, and the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 1.* Cambridge, MA: MIT Press.

Schneider, W. (1985). Toward a model of attention and the development of automatic processing. In M.I. Posner & O.S.M. Marin (Eds.), *Attention and performance XI* (pp.475-492). Hillsdale, NJ: Lawrence Erlbaum.

Schneider, W. & Detweiler, M. (1987). A connectionist/control architecture for working memory. In G. Bower (Ed.), *The psychology of learning and motivation, Vol. 21.* New York: Academic Press.

Schneider, W. & Oliver, W. L. (in press). An instructable connectionist/control architecture: Using rule-based instructions to accomplish connectionist learning in a human time scale.

Schneider, W. & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review, 84,* 1-66.

Shaffer, W. O. (1975). Multiple attention in continuous verbal tasks. In P.M.A. Rabbitt and S. Dornic (Eds.), *Attention and performance V.* New York: Academic Press.

Shiffrin, R. M. (in press) Attention. To appear in R.C. Atkinson, R.J. Herrnstein, G. Lindzey, and R.D. Luce (Eds.), *Steven's handbook of experimental psychology, 2nd edition.* New York: John Wiley & Sons, Inc.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review, 84,* 127-190.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18,* 643-662.

Treisman, A. M. (1960). Contextual cues in selective listening. *Quarterly Journal of Experimental Psychology, 12,* 242-248.

Wickens, D. D. (1984). Processing resources in attention. In R. Parasuraman & D. R. Davies (Eds.), *Varieties of attention.* Orlando, FL: Academic Press.

Appendix A: Parameters of the Model

Performance of the model depended upon a number of different parameters. Several of these were tightly constrained by the data, while the values of others seemed to be less critical. Here, we review all of the parameters that were relevant to the simulations reported, and provide a rationale for how the value of each was chosen, how critical that value was and, where appropriate, which other parameters it interacted with.

Ratio of training frequencies: The network used in the first two simulations was trained on words and colors in a 10:1 ratio. This value was chosen to capture both the difference in the speed of processing between word reading and color naming and the size of the interference and facilitation effects observed for color naming (see Figure 5a). This parameter interacted primarily with the magnitude of the attentional influence in the model, the parameters of the response mechanism, and the maximum response time (see descriptions below). The actual value of this parameter did not appear to be critical, and values ranging from 5:1 to 20:1 gave comparable results, providing the size of the parameters mentioned above were adjusted to compensate. The actual value was not considered to be crucial, but the asymmetry in training that it represented is theoretically important: it is differential amounts of training that lead to differential pathway strengths. This is consistent with the common assumption that word reading is more highly practiced than color naming.

Learning rate: This parameter scaled the size of the changes made to connection weights in each learning trial. Its value was tightly constrained by the MacLeod and Dunbar (1988) data. The exact same number of training trials per stimulus were used in the simulation as were used in the empirical study. We chose a learning rate that — given this number of trials — produced the closest fit to the interference data at each test point. This parameter was the same for all of the connections in each direct pathway in the network (including the input connections). However, the connections in the indirect pathway used in Simulation 4 were fixed; that is, they had a learning rate of 0.

Maximum response time: This parameter functioned as our training criterion for Simulation 1. After specifying a learning rate (see above), we needed some way of

deciding when to stop training on colors and words. Training ended when the network could respond accurately to all of the test stimuli (control, conflict and congruent stimuli in each task) within a specified number of cycles, which we call the maximum response time. A lower value (faster response) meant more training, and a higher value meant less training. The value we used was 50 cycles. There were two primary constraints on this value: a) test performance after training had to simulate the basic Stroop effect (see Figure 5); b) regression of simulation cycles on empirical reaction time data had to yield a positive intercept value of reasonable magnitude. A small or zero valued intercept would suggest — unreasonably — that our model simulates *all* of the processes involved in human performance; a negative intercept would indicate that the effects of interest were only a small part of the model's overall performance. The intercept values for all simulations reported were in the 200-400 msec range. Both of these constraints on the maximum reaction time were relatively weak, and reasonable performance was achievable with a broad range of values. This parameter interacted with the training frequency ratio and the noise parameters in the network (see below).

Preset input weights: The weights from the input to the intermediate units in each pathway were given preassigned values at the outset of training. This was required for changes in reaction time with training to follow a power law. This finding is theoretically important, for it suggests that the power law may only apply to the learning of simple mappings, and that learning at the early stages of processing (e.g., stimulus encoding) are not involved in the tasks we have simulated. The actual values preassigned to the input weights were picked based on the weights that are achieved when the network is allowed to learn this set of weights on its own.

Indirect pathway: This was the module and corresponding set of weights that was added to the basic network for Simulation 4. The number of units in this module was the same as in all others (2). The strengths assigned to the connections in this pathway were chosen to fit best the empirical data concerning reaction times and interference effects early in training on shape naming.

Parameters of the response mechanism: Three parameters were associated with the response mechanism: the rate at which evidence accumulated, the noise associated with this process, and the threshold for a response. The rate and threshold parameters were inversely related to one another: doubling the accumulation rate was equivalent to halving the response threshold. Together these values interacted with the maximum response time

to determine the networks performance for a given amount of training. They also showed a complex interaction with the cascade rate (see below), that affected the relative magnitude of interference effects versus speed of processing differences between the pathways. Values for these parameters were chosen which provided the best fit to the basic Stroop effect, given the constraints imposed by the other parameters of the models (i.e., learning rate, maximum response time, and attentional influence). The amount of noise associated with the response mechanism interacted with the amount of noise added to the net input of processing units. The constraints on both noise parameters are discussed below.

Cascade rate: This parameter determined the rate at which each unit accumulated activation. Its value was the same for all units in the network (except input units, which were always maximally excited or inhibited). The cascade rate interacted with the response rate and response threshold to determine both reaction time and the pattern of interference effects between processes. Values for these parameters were chosen which provided the best fit to the basic Stroop effect, given the constraints imposed by the other parameters of the models (i.e., learning rate, maximum response time, and attention influence).

Noise: This scaled the magnitude of Gaussian distributed noise added to the net input of each unit (except the input units). Values of this parameter and the noise in the response mechanism were chosen to provide maximum variance without sacrificing accuracy of performance. This interacted with the maximum response time, to determine the amount of training received by the networks used in Simulations 1,2, 5 and 6. These parameters also had an influence on the relationship between mean reaction time and variance as a function of training. Values were chosen — within the limits discussed above — which provided the closest match between the exponents of the power functions describing the changes in mean reaction time and standard deviation with practice.

Magnitude of attentional influence: Two parameters determined the magnitude of attentional effects in the model. These were the size of the resting negative bias on intermediate units in the two processing pathways, and the size of the weights from the task demand units to these intermediate units. The magnitudes of these two parameters were constrained to be equal, so that with full activation of a given task demand unit, intermediate units in the corresponding pathway had a resting net input of zero and an activation of 0.5 (the rationale for this is explained in Attentional Selection under The Model). The main effect of varying these parameters, therefore, was to change the resting activation level of intermediate units in the *unattended* pathway (since their task demand

unit was not active, and therefore their negative bias was not offset). The magnitude of the attentional influence interacted with the amount of training received by the network, influencing the size of the interference and facilitation effects observed. A set of values was chosen that provided the best fit to the empirical data in each experiment (see "Free parameters" under Simulation Methods).

Number of units/module: This was essentially a free parameter. In order to keep the networks as simple as possible, we chose the fewest number of units possible to simulate each task.